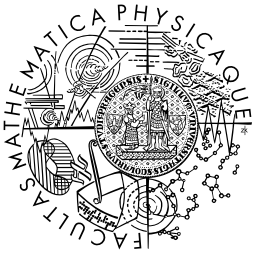

MATEMATICKO-FYZIKÁLNÍ FAKULTA
Univerzita Karlova

DIPLOMOVÁ PRÁCE

Martina Sotáková

Zobecněné odhadovací rovnice (GEE)

Katedra pravděpodobnosti a matematické statistiky

Vedoucí diplomové práce:	doc. Ing. Marek Omelka, Ph.D.
Studijní program:	Matematika
Studijní obor:	Finanční a pojistná matematika

Praha 2020

Prohlašuji, že jsem tuto diplomovou práci vypracoval(a) samostatně a výhradně s použitím citovaných pramenů, literatury a dalších odborných zdrojů.

Beru na vědomí, že se na moji práci vztahují práva a povinnosti vyplývající ze zákona č. 121/2000 Sb., autorského zákona v platném znění, zejména skutečnost, že Univerzita Karlova má právo na uzavření licenční smlouvy o užití této práce jako školního díla podle §60 odst. 1 autorského zákona.

V Prahe dne 30.7.2020 Martina Sotáková

Na tomto mieste by som chcela poďakovať doc. Ing. Marekovi Omelkovi, Ph.D., za obetovaný čas pri vedení práce, trpezlivosť a cenné rady.
Poďakovanie patrí aj mojej rodine a blízkym priateľom za prejavenú podporu počas celého štúdia.

Název práce: Zobecněné odhadovací rovnice (GEE)

Autor: Martina Sotáková

Katedra: Katedra pravděpodobnosti a matematické statistiky

Vedoucí diplomové práce: doc. Ing. Marek Omelka, Ph.D., Katedra pravděpodobnosti a matematické statistiky

Abstrakt: V tejto práci sa zaoberáme zovšeobecnenými odhadovacími rovnicami (GEE). Najskôr zavádzame pojem zovšeobecneného lineárneho modelu, na ktorom sú zovšeobecnené odhadovacie rovnice postavené. Ďalej sú predstavené metódy pseudo maximálnej vierohodnosti a kvázi pseudo maximálnej vierohodnosti, z ktorých prechádzame k metóde zovšeobecnených odhadovacích rovníc. Na záver sú prevedené simulačné štúdie, ktoré demonštrujú teoretické výsledky uvedené v práci.

Klíčová slova: zovšeobecnený lineárny model, zovšeobecnené odhadovacie rovnice, pseudo maximálna vierohodnosť, kvázi pseudo maximálna vierohodnosť, QIC

Title: Generalized estimating equations

Author: Martina Sotáková

Department: Department of Probability and Mathematical Statistics

Supervisor: doc. Ing. Marek Omelka, Ph.D., Department of Probability and Mathematical Statistics

Abstract: In this thesis we are interested in generalized estimating equations (GEE). First, we introduce the term of generalized linear model, on which generalized estimating equations are based. Next we present the methos of pseudo maximum likelyhood and quasi-pseudo maximum likelyhood, from which we move on to the methods of generalized estimating equations. Finally, we perform simulation studies, which demonstrates the theoretical results presented in the thesis.

Keywords: generalized linear model, generalized estimating equations, pseudo maximum likelihood, quasi pseudo maximum likelihood, QIC

Obsah

Úvod

Zovšeobecnený lineárny model je nástroj, ktorý je pomerne často využívaný v oblasti poisťovníctva. Už od jeho predstavenia autormi Nelder a Wedderburn (1972) sa stal populárnym, pretože rozdelenie vysvetľovanej premennej sa neobmedzuje len na normálne rozdelenie, ale rozsah rozdelení je rozšírený o rozdelenia patriace do rodiny exponenciálnych rozdelení. Zväčšenie tohto rozsahu nám umožňuje pracovať aj s diskrétnymi dátami.

Pri modelovaní pomocou zovšeobecneného lineárneho modelu ale predpokladáme, že všetky pozorovania sú navzájom nezávislé. Bohužiaľ, tento predpoklad môže byť často krát porušený, v niektorých prípadoch až nereálny. Preto bolo potrebné nájsť metódu, ktorá bude brať do úvahy závislosť medzi pozorovaniami. Liang a Zeger (1986) vo svojom článku opisujú metódu zovšeobecnených odhadovacích rovníc (GEE) pre odhad parametrov modelu, kde je predpoklad nezávislosti porušený. Túto metódu odvodili pre takzvané skupinovo závislé dáta - pozorovania, ktoré sú závislé v skupine, ale nezávislé naprieč skupinami.

Táto práca sa venuje metóde zovšeobecnených odhadovacích rovníc (GEE). Cieľom práce je čitateľovi predstaviť túto metódu a na vhodných simuláciách demonštrovať teoretické výsledky.

Práca je delená do troch častí. Prvá časť práce je venovaná zovšeobecneným lineárnym modelom. Ako prvú uvádzame rodinu exponenciálnych rozdelení. Ďalej formulujeme zovšeobecnený lineárny model a ukážeme, ako nájsť odhady parametrov v tomto modeli metódou maximálnej vierohodnosti.

Druhá časť práce sa venuje metódam zovšeobecnených odhadovacích rovníc. Postupne predstavíme metódy, na základe ktorých sa postupne dostaneme k metóde zovšeobecnených odhadovacích rovníc (GEE). Predstavíme metódu pseudo maximálnej vierohodnosti, metódu kvázi psuedo maximálnej vierohodnosti až sa nakoniec dostaneme k samotnej metóde GEE. Na príkladoch ukážeme využitie týchto metód.

V tretej časti práce na vhodne zvolených simuláciách ukážeme, ako vplýva počet skupín na vlastnosti odhadu parametru v modeli a ako vhodne vybrať pracovnú korelačnú štruktúru pomocou kvázi Akaikovho kritéria (QIC).

1. Úvod do zovšeobecnených lineárnych modelov

V tejto časti si najprv predstavíme pojem zovšeobecneného lineárneho modelu (generalized linear model, GLM), na ktorom neskôr postavíme teóriu zovšeobecnených odhadovacích rovníc. Zdrojom, o ktorý sa budeme opierať, je Kulich (2020). Predstavíme predpoklady tohto modelu a vlastnosti odhadnutých parametrov. Zovšeobecnený lineárny model je rozšírením klasického lineárneho modelu, ktoré špecifikuje strednú hodnotu odozvy ako nejakú funkciu lineárnej kombinácie vysvetľujúcich premenných. Pre lineárny zovšeobecnený model máme väčšiu škálu pre výber rozdelenia odozvy. Tieto rozdelenia pochádzajú z takzvanej rodiny exponenciálnych rozdelení.

1.1 Rodina exponenciálnych rozdelení

Tvar hustoty rozdelení, ktoré pochádzajú z rodiny exponenciálnych rozdelení (the exponential family of distributions), vieme vyjadriť ako

$$f(x;\theta,\varphi) = \exp\left\{\frac{x\theta - b(\theta)}{\varphi} + c(x,\varphi)\right\}, x \in M \tag{1.1}$$

vzhľadom k nejakej σ-konečnej miere μ. Množina M je nosičom pre také x, pre ktoré je hustota kladná. Funkcie $b(.)$ a $c(.)$ sú reálne funkcie. Tieto funkcie sa líšia pre každé rozdelenie náležiace do rodiny exponenciálnych rozdelení. Parameter $\theta \in \mathbb{R}$ nazývame kanonickým parametrom a parameter $\varphi \in (0,\infty)$ nazývame disperzným parametrom.

Poznámka.

1. Za σ-konečnú mieru μ budeme uvažovať Lebesguovu mieru alebo čítaciu mieru.

2. Výraz (1.1) sa nazýva kanonický tvar hustoty.

Do tejto skupiny rozdelení patria napríkald normálne rozdelenie $\mathrm{N}(\mu,\sigma^2)$, Poissonovo rozdelenie $\mathrm{Po}(\lambda)$, gamma rozdelenie $\Gamma(a,p)$, alternatívne rozdelenie $\mathrm{Alt}(p)$ a iné. Môžeme si všimnúť, že rodina exponenciálnych rozdelení združuje ako spojité, tak aj diskrétne rozdelenia.

Príklad. Ukážeme, že Poissonovo rozdelenie patrí do rodiny exponenciálnych rozdelení. Majme náhodnú veličinu X, ktorá sa riadi Poissonovým rozdelením vzhľadom k čítacej miere μ s hustotou

$$f(x;\lambda) = \frac{\lambda^x}{x!}\exp\{-\lambda\}, \quad \lambda > 0,\ x \in \mathbb{N}_0.$$

Úpravou dostaneme tvar

$$f(x;\lambda) = \exp\left\{-\lambda + \log\left(\frac{\lambda^x}{x!}\right)\right\} = \exp\{x\ \log\ \lambda - \lambda + \log x!\}.$$

Nasledujúcou substitúciou dostaneme tvar hustoty v (1.1)

$$\theta = \log\lambda, \quad \varphi = 1, \quad b(\theta) = \exp(\theta), \quad c(x,\varphi) = \log x!.$$

Príklad. V tomto príklade ukážeme, že normálne rozdelenie taktiež patrí do rodiny exponenciálnych rozdelení. Majme náhodnú veličinu X, ktorá sa riadi normálnym rozdelením so strednou hodnotou μ a rozptylom σ^2. Toto rozdelenie má hustotu v tvare

$$f(x;\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(\frac{-(x-\mu)^2}{2\sigma^2}\right), \quad \mu\in\mathbb{R},\ \sigma^2>0,\ x\in\mathbb{R}.$$

Úpravami dostávame tvar

$$f(x;\mu,\sigma^2) = \exp\left\{\frac{\mu x}{\sigma^2} - \frac{1}{2}\log(2\pi\sigma^2) - \frac{x^2}{2\sigma^2} - \frac{\mu^2}{2\sigma^2}\right\}.$$

Substitúciou dostávame tvar hustoty v (1.1)

$$\theta = \mu, \quad \varphi = \sigma^2, \quad b(\theta) = \frac{\theta^2}{2}, \quad c(x,\varphi) = -\frac{x^2}{2\varphi} - \frac{1}{2}\log(2\pi\varphi).$$

Pre náhodné veličiny, ktorých rozdelenie patrí do rodiny exponenciálnych rozdelení, môžeme explicitne vyjadriť predpis pre momentovú vytvárajúcu funkciu pomocou následujúcej lemmy:

Lemma 1. *Nech sa náhodná veličina X riadi rozdelením, ktoré patrí do rodiny exponenciálnych rozdelení. Potom momentová vytvárajúca funkcia $M_X(t) \equiv \mathbb{E}e^{tX}$ náhodnej veličiny X existuje, je konečná a je rovná výrazu*

$$M_X(t) = \exp\left\{\frac{b(t\varphi+\theta) - b(\theta)}{\varphi}\right\}.$$

Dôkaz. Označme M množinu, na ktorej je hustota náhodnej veličiny X vzhľadom ku σ-konečnej miere μ kladná.

Z definície momentovej vytvárajúcej funkcie môžeme počítať

$$\begin{aligned}
M_X(t) &= \int_M \exp\{tx\}\exp\left\{\frac{x\theta - b(\theta)}{\varphi} + c(x,\varphi)\right\} d\mu(x)\\
&= \int_M \exp\left\{\frac{x(t\varphi+\theta) - b(\theta)}{\varphi} + c(x,\varphi)\right\} d\mu(x)\\
&= \int_M \exp\left\{\frac{x(t\varphi+\theta) - b(\theta) + b(t\varphi+\theta) - b(t\varphi+\theta)}{\varphi} + c(x,\varphi)\right\} d\mu(x)\\
&= \exp\left\{\frac{b(t\varphi+\theta) - b(\theta)}{\varphi}\right\}\int_M \exp\left\{\frac{x(t\varphi+\theta) - b(t\varphi+\theta)}{\varphi} + c(x,\varphi)\right\} d\mu(x).
\end{aligned}$$

V poslednej rovnosti vidíme, že integrujeme funkciu, ktorá má tvar hustoty pochádzajúcej z rodiny exponenciálnych rozdelení (využijeme vhodne zvolenú substitúciu kanonického parametru). Tento integrál je z vlastností hustoty rovný jednej a dostávame sa k výrazu, ktorý sme hľadali.

□

Ak je funkcia $b(\theta)$ dvakrát spojito diferencovateľná, potom je $M_X(t)$ taktiež dvakrát diferencovateľná v bode $t = 0$ a platí $\mathbb{E}X = b'(\theta)$ a var $X = \varphi\, b''(\theta)$. To ukážeme na základe výsledku vyššie uvedenej lemmy a z vlastností momentovej vytvárajúcej funkcie. Vieme, že platí vzťah (viď Zvára a Štepán (1997), Veta 7.7)

$$\mathbb{E}X^n = M_X^{(n)}(t)|_{t=0}.$$

Prvou a druhou deriváciou vyššie odvodeného výrazu dostávame

$$\begin{aligned} M'_X(t) &= \exp\left\{\frac{b(t\varphi+\theta)-b(\theta)}{\varphi}\right\} b'(t\varphi+\theta) \\ M''_X(t) &= \exp\left\{\frac{b(t\varphi+\theta)-b(\theta)}{\varphi}\right\} [b'(t\varphi+\theta)]^2 \\ &\quad + \exp\left\{\frac{b(t\varphi+\theta)-b(\theta)}{\varphi}\right\} b''(t\varphi+\theta)\varphi. \end{aligned}$$

Dosadením $t = 0$ vieme vyjadriť strednú hodnotu a rozptyl náhodnej veličiny X ako

$$\begin{aligned} \mathbb{E}X &= M'_X(0) = b'(\theta) \\ var\, X &= \mathbb{E}X^2 - (\mathbb{E}X)^2 = M''_X(0) - [M'_X(0)]^2 \\ &= [b'(\theta)]^2 + \varphi\, b''(\theta) - [b'(\theta)]^2 = \varphi\, b''(\theta). \end{aligned}$$

Označme $\mu = \mathbb{E}X = b'(\theta)$. Ďalej budeme predpokladať, že funkcia $b(\theta)$ bude stále dvakrát spojito diferencovateľná. Z tohoto predpokladu je druhá derivácia funkcie $b(.)$ konečná, čím máme zaručené, že rozptyl nedegenerovanej náhodnej veličiny X je taktiež konečný. Keďže jednou z vlastností rozptylu nedegenerovanej náhodnej veličiny je, že je kladný, tak z tejto vlastnosti vidíme, že funkcia $b'(\theta)$ je rýdzo rastúcou funkciou. Odtiaľ vieme, že funkcia $b'(\theta)$ bude mať správne zadefinovanú inverznú funkciu $(b')^{-1}(\theta)$. Toto neskôr využijeme pri parametrizácii zovšeobecneného lineárneho modelu.

Definícia 1. *Nech existuje funkcia $V(\mu)$ taká, že platí var $X = \varphi\, V(\mu)$ a platí $b''(\theta) = V(b'(\theta))$. Takúto funckiu $V(\mu)$ nazývame* rozptylová funkcia.

Vidíme, že rozptylová funkcia ukazuje závislosť rozptylu náhodnej veličiny na jej strednej hodnote. Ak rozdelenie pochádza z rodiny exponenciálnych rozdelení, potom rozptylová funkcia jednoznačne určuje rozdelenie náhodnej veličiny.

1.1.1 Mnohorozmerná rodina exponenciálnych rozdelení

Pre ďalšie použitie, v druhej kapitole tejto práce, ukážeme, ako vyzerá tvar hustoty exponenciálnej rodiny rozdelení v mnohorozmernom prípade. Uvažujme

$\boldsymbol{X} \in \mathbb{R}^n$ n-rozmerný náhodný vektor, $\boldsymbol{\theta} \in \Theta \subset \mathbb{R}^n$ vektor kanonického parametra a $\boldsymbol{\Phi} \in \mathbb{R}^{n\times n}$ maticu disperzného parametru. Tvar hustoty n-dimenzionálneho rozdelenia patriaceho do n-dimenzionálnej rodiny exponenciálnych rozdelení je tvaru

$$f(\boldsymbol{x},\boldsymbol{\theta},\boldsymbol{\Phi}) = \exp\left(\boldsymbol{\theta}^\top \boldsymbol{x} - b(\boldsymbol{\theta},\boldsymbol{\Phi}) + c(\boldsymbol{x},\boldsymbol{\Phi})\right), \tag{1.2}$$

kde b a c sú funkcie z $\mathbb{R}^n \times \mathbb{R}^{n\times n}$ do $\mathbb{R}$.

Príklad. Uvažujme, že náhodný vektor $\boldsymbol{X} \in \mathbb{R}^n$ sa riadi mnohorozmerným normálnym rozdelením so strednou hodnotou $\boldsymbol{\mu}$ a kovariančnou maticou $\boldsymbol{\Sigma}$. Toto rozdelenie má tvar hustoty

$$f(\boldsymbol{x};\boldsymbol{\mu},\boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{\frac{n}{2}}\det(\boldsymbol{\Sigma})^{\frac{1}{2}}} \exp\left\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right\}, \quad \boldsymbol{x} \in \mathbb{R}^n.$$

Úpravami sa dostaneme na tvar

$$f(\boldsymbol{x};\boldsymbol{\mu},\boldsymbol{\Sigma}) = \exp\left\{\boldsymbol{\theta}^\top \boldsymbol{x} - \frac{n}{2}\ln(2\pi) - \frac{1}{2}\boldsymbol{x}^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{x} - \frac{1}{2}\ln(\det(\boldsymbol{\Sigma})) - \frac{1}{2}\boldsymbol{\theta}^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\theta}\right\}$$

a substitúciou

$$\boldsymbol{\theta} = \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}, \quad b(\boldsymbol{\theta},\boldsymbol{\Sigma}) = \frac{1}{2}\boldsymbol{\theta}^\top \boldsymbol{\Sigma}\boldsymbol{\theta},$$
$$c(\boldsymbol{x},\boldsymbol{\Sigma}) = -\frac{n}{2}\ln(2\pi) - \frac{1}{2}\boldsymbol{x}^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{x} - \frac{1}{2}\ln(det(\boldsymbol{\Sigma}))$$

dostávame tvar hustoty uvedený v (1.2).

Tak, ako v prípade jednorozmernej rodiny exponenciálnych rozdelení, vieme vyjadriť vzťah aj pre momenty náhodného vektoru $\boldsymbol{X}$, ktorý sa riadi rozdelením s hustotou v tvare (1.2) (viď Ziegler (2011), Theorem 1.2.):

$$\mathbb{E}\boldsymbol{X} = \boldsymbol{\mu} = \frac{\partial b(\boldsymbol{\theta},\boldsymbol{\Phi})}{\partial \boldsymbol{\theta}},$$
$$var\ \boldsymbol{X} = \boldsymbol{\Sigma} = \frac{\partial^2 b(\boldsymbol{\theta},\boldsymbol{\Phi})}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}^\top}.$$

Pre určité prípady je vhodné, pokiaľ je (1.2) vyjadrená v závislosti na $\boldsymbol{\mu}$ namiesto závislosti na $\boldsymbol{\theta}$. Tým sa dostaneme na tvar

$$f(\boldsymbol{x},\boldsymbol{\mu},\boldsymbol{\Phi}) = \exp\left(a(\boldsymbol{\mu},\boldsymbol{\Phi})^\top \boldsymbol{x} + d(\boldsymbol{\mu},\boldsymbol{\Phi}) + c(\boldsymbol{x},\boldsymbol{\Phi})\right), \tag{1.3}$$

kde $a(\boldsymbol{\mu},\boldsymbol{\Phi}) = \boldsymbol{\theta}$ a $d(\boldsymbol{\mu},\boldsymbol{\Phi}) = -b(a(\boldsymbol{\mu},\boldsymbol{\Phi}),\boldsymbol{\Phi})$.

Vybrané vlastnosti n-dimenzionálnej rodiny exponenciálnych rozdelení, ktoré v následujúcich častiach práce využijeme, môžeme nájsť v prílohe A.1

Poznámka. Pre $n = 1$ má hustota v (1.2) tvar

$$f(x;\theta_2,\Phi) = \exp\left\{\theta_2 x - b_2(\theta_2,\Phi) + c_2(x,\Phi)\right\},$$

kde $\theta_2 \in \mathbb{R}$ je kanonický parameter, $\Phi \in (0,\infty)$ je disperzný parameter, funkcie b_2 a c_2 sú funkcie z $\mathbb{R} \times \mathbb{R}$ do $\mathbb{R}$. Pre prehľadnosť vzťahov sme použili značenie s dolným indexom 2. Pre porovnanie prepíšeme hustotu z (1.1)

$$f(x;\theta_1,\varphi) = \exp\left\{\frac{x\theta_1 - b_1(\theta_1)}{\varphi} + c_1(x,\varphi)\right\}.$$

Odtiaľ hneď vidíme, že disperzný parameter $\varphi = \Phi$. Ďalej môžeme vidieť, že platia nižšie uvedené vzťahy

$$\theta_1 = \varphi \cdot \theta_2, \quad b_1(\theta_1) = \varphi \cdot b_2(\theta_2,\Phi), \quad c_1(x,\varphi) = c_2(x,\Phi).$$

1.2 Zovšeobecnený lineárny model

Pojem zovšeobecneného lineárneho modelu bol prvýkrát sformulovaný v článku Nelder a Wedderburn (1972) a zjednotil rôzne regresné metódy ako napríklad lineárnu regresiu, logistickú regresiu a iné.

Na začiatok si zavedieme značenie. Budeme uvažovať n nezávislých pozorovaní náhodného vektoru $(Y_i,\boldsymbol{X}_i), i = 1,\ldots,n$, kde $\boldsymbol{X}_i = (X_{i1},\ldots,X_{ip})^\top$. Ďalej označme $\boldsymbol{\varepsilon} = (\varepsilon_1,\ldots,\varepsilon_n)^\top$ ako vektor chýb. V zovšeobecnenom lineárnom modeli platí

$$Y_i = \mu_i + \varepsilon_i, \tag{1.4}$$

kde $\mu_i = \mathbb{E}[Y_i|\boldsymbol{X}_i]$ je podmienená stredná hodnota Y_i za podmienky $\boldsymbol{X}_i$.

Tak, ako v klasickom lineárnom modeli, aj v zovšeobecnenom lineárnom modeli chceme vyjadriť vzťah medzi $\mu_i = \mathbb{E}[Y_i|\boldsymbol{X}_i]$ a $\boldsymbol{X}_i$. Narozdiel od klasického modelu tento vzťah nemusí byť len lineárny. V zovšeobecnenom lineárnom modeli chceme túto závislosť vyjadriť v širšom zábere, ako je to v prípade klasického lineárneho modelu.

Pre formuláciu zovšeobecneného lineárneho modelu budeme musieť špecifikovať:

- podmienené rozdelenie pozorovaní $Y_i|\boldsymbol{X}_i$,
- lineárny prediktor,
- linkovú funkciu.

Začneme špecifikáciou podmieneného rozdelenia $Y_i|\boldsymbol{X}_i$. Ako už bolo povedané, zovšeobecnený lineárny model je rozšírením klasického lineárneho modelu. Pre podmienené rozdelenie Y_i za podmienky $\boldsymbol{X}_i$ sa teda nemusíme obmedzovať len na normálne rozdelenie, ale máme široký výber z rozdelení pochádzajúcich z rodiny exponenciálych rozdelení, o ktorej sme sa už v tejto práci zmienili v časti 1.1. Hustota má pre tieto rozdelenia tvar ako v (1.1), kde $b(.)$ je dvakrát diferencovateľná funkcia a parameter θ_i závisí na lineárnej kombinácii $\boldsymbol{X}_i$ a $\boldsymbol{\beta}$. Taktiež pozorovania Y_i sú závislé na lineárnej kombinácii $\boldsymbol{X}_i$ a $\boldsymbol{\beta}$.

Lineárny prediktor η_i je jednoducho lineárna kombinácia regresných koeficientov $\boldsymbol{\beta}$ a vysvetľujúcich premenných $\boldsymbol{X}_i$

$$\eta_i = \beta_1 X_{i1} + \cdots + \beta_p X_{ip} = \boldsymbol{X}_i^\top \boldsymbol{\beta}.$$

Názov lineárny má z dôvodu, že η_i je lineárny vzhľadom k regresným koeficientom.

Nakoniec, linková funkcia špecifikuje vzťah medzi strednou hodnotou μ_i a lineárnym prediktorom η_i. Linkovou funkciou nazveme takú funkciu g, ktorá je rýdzo monotónna, dvakrát diferencovateľná a spĺňuje

$$g(\mu_i) = \eta_i = \boldsymbol{X}_i^\top \boldsymbol{\beta}.$$

Linková funkcia musí byť stále určená vopred. V tabuľke 1.1 môžeme vidieť najčastejšie používané linkové funkcie.

Názov linkovej funkcie	**Linková funkcia** $g(\mu) = \eta$
Identita	μ
Log	$\log \mu$
Log-log	$-\log(-\log \mu)$
Logit	$\log(\frac{\mu}{1-\mu})$
Probit	$\Phi^{-1}(\mu)$

Tabuľka 1.1: Linkové funkcie.

Poznámka. Ak linková funkcia g spĺňuje vzťah $g(\mu) = \theta$, potom o takejto linkovej funkcii hovoríme ako o *kanonickej linkovej funkcii*. Každé rozdelenie, ktoré patrí do rodiny exponenciálnych rozdelení, má jednoznačne určenú kanonickú linkovú funkciu.

Primárnym cieľom GLM je odhad regresných koeficientov $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^\top$. Zo špecifikácie zovšeobecneného modelu môžeme vidieť, že parametre modelu sú závislé na týchto regresných koeficientoch. Keďže tieto parametre parametrizujú aj rozdelenie Y_i, budeme ich chcieť taktiež odhadnúť. Parametre rozdelenia Y_i teda môžeme parametrizovať troma spôsobmi: pomocou lineárneho prediktoru η_i, strednej hodnoty μ_i alebo kanonického parametru θ_i príslušného pre jednotlivé pozorovania Y_i. Rovnosti v (1.5) udávajú vzťahy medzi týmito parametrami:

$$\begin{aligned} &\eta_i = g(\mu_i), && \mu_i = g^{-1}(\eta_i) \\ &\mu_i = b'(\theta_i), && \theta_i = (b')^{-1}(\mu_i) \\ &\eta_i = g(b'(\theta_i)), && \theta_i = (b')^{-1}(g^{-1}(\eta_i)). \end{aligned} \tag{1.5}$$

Tieto vzťahy medzi jednotlivými parametrami využijeme neskôr pri hľadaní odhadu regresných koeficientov.

1.3 MLE odhady parametrov v GLM

Najčastejšou metódou, ktorou odhadujeme regresné koeficienty, býva metóda maximálnej vierohodnosti (maximum likelihood method, ML). Táto metóda je

veľmi často využívaná aj v prípade GLM modelu. Uvažujme n nezávislých, rovnako rozdelených pozorovaní $(Y_i,\boldsymbol{X}_i)^\top$, ktoré splňujú predpoklady GLM modelu. Ďalej predpokladajme, že marginálne rozdelenie $\boldsymbol{X}_i$ sa riadi rozdelením s hustotou h, ktorá nie je závislá na parametroch $\boldsymbol{\beta}$ a φ. Vierohodnostnú funkciu pre regresné koeficienty $\boldsymbol{\beta}$ môžeme zapísať v tvare

$$\begin{aligned} L(\boldsymbol{\beta}|\boldsymbol{Y}) &= \prod_{i=1}^{n} f(Y_i; \boldsymbol{X}_i, \boldsymbol{\beta},\varphi) \\ &= \prod_{i=1}^{n} f(Y_i|\boldsymbol{X}_i; \boldsymbol{\beta},\varphi) h(\boldsymbol{X}_i) \\ &= \prod_{i=1}^{n} \exp\left\{ \frac{Y_i\theta_i - b(\theta_i)}{\varphi} + c(Y_i,\varphi) \right\} h(\boldsymbol{X}_i), \end{aligned} \tag{1.6}$$

kde θ_i je parametrizovaná ako v (1.5) a kde sme v druhej rovnosti použili vzťah medzi marginálnou hustotou a podmienenou hustotou (viď Lachout (2004), Veta 9.6).

Hľadáme také $\boldsymbol{\beta}$, ktoré maximalizuje hodnotu vierohodnostnej funkcie (1.6). Obvykle sa vierohodnostná funkcia upraví do tvaru logaritmickej vierohodnostnej funkcie, čo je v tomto prípade funkcia tvaru

$$\begin{aligned} \ell(\boldsymbol{\beta}|\boldsymbol{Y}) &= \log L(\boldsymbol{\beta}|\boldsymbol{Y}) \\ &= \sum_{i=1}^{n} \left[\frac{Y_i\theta_i - b(\theta_i)}{\varphi} + c(Y_i,\varphi) \right] + \sum_{i=1}^{n} \log h(\boldsymbol{X}_i). \end{aligned} \tag{1.7}$$

Vidíme, že posledný člen (1.7) nezávisí na koeficientoch $\boldsymbol{\beta}$ a preto tento člen nebude hrať úlohu pri maximalizácii tejto funkcie. Taktiež členy $c(Y_i,\varphi)$ nezávisia na parametroch $\boldsymbol{\beta}$, teda s nimi taktiež nemusíme počítať pri maximalizácii (1.7). V konečnom dôsledku budeme maximalizovať upravenú logaritmickú vierohodnostnú funkciu

$$\ell^*(\boldsymbol{\beta}|\boldsymbol{Y}) = \sum_{i=1}^{n} \frac{Y_i\theta_i - b(\theta_i)}{\varphi}. \tag{1.8}$$

K nájdeniu vektoru $\boldsymbol{\beta}$, ktorý maximalizuje upravenú logaritmickú vierohodnostnú funkciu, zderivujeme výraz v (1.8) vzhľadom k zložkám β_j s využitím retiazkového pravidla a vzťahov, ktoré sú uvedené v (1.5). Skórová funkcia je tvaru

$$\boldsymbol{U}(\boldsymbol{\beta}) = \left[\frac{\partial \ell^*}{\partial \beta_j} \right]_{j=1,\ldots,p} = \left[\sum_{i=1}^{n} \frac{\partial \ell^*}{\partial \theta_i} \frac{\partial \theta_i}{\partial \mu_i} \frac{\partial \mu_i}{\partial \eta_i} \frac{\partial \eta_i}{\partial \beta_j} \right]_{j=1,\ldots,p}. \tag{1.9}$$

Jednotlivé derivácie vo výraze (1.9) sú rovné

$$\begin{aligned} \frac{\partial \ell^*}{\partial \theta_i} &= \frac{Y_i - b'(\theta_i)}{\varphi} = \frac{Y_i - \mu_i}{\varphi}, \\ \frac{\partial \theta_i}{\partial \mu_i} &= \frac{1}{b''((b')^{-1}(\mu_i))} = \frac{1}{b''(\theta_i)} = \frac{1}{V(\mu_i)}, \\ \frac{\partial \mu_i}{\partial \eta_i} &= \frac{1}{g'(g^{-1}(\eta_i))} = \frac{1}{g'(\mu_i)}, \\ \frac{\partial \eta_i}{\partial \beta_j} &= X_{ij}, \end{aligned}$$

kde sme využili deriváciu inverznej funkcie a vzťahy uvedené v (1.5). Pre nájdenie odhadu $\hat{\boldsymbol{\beta}}$ položíme skórovú funkciu rovnú nulovému p-rozmernému vektoru, teda platí

$$\boldsymbol{U}(\boldsymbol{\beta}) = \left[\sum_{i=1}^{n} \frac{Y_i - \mu_i}{\varphi} \frac{1}{V(\mu_i)} \frac{1}{g'(\mu_i)} X_{ij}\right]_{j=1,\ldots,p} = \mathbf{0}_{p\times 1}. \tag{1.10}$$

Odhad $\widehat{\boldsymbol{\beta}}$ regresných koeficientov $\boldsymbol{\beta}$ bude riešením rovnice (1.10). Tento odhad nazývame maximálne vierohodným odhadom (maximum likelihood estimator, MLE). Rovnica (1.10) sa rieši numericky za pomoci iteračných metód. Medzi najpoužívanejšie iteračné metódy pre tento problém patria napríklad metóda iterovaných vážených najmenších štvorcov či Newtonova- Raphsonova metóda. Algoritmy k obom metódam môžeme nájsť v Hardin a Hilbe (2003), kapitola 2.

2. Zovšeobecnené odhadovacie rovnice

Táto kapitola bude pojednávať o metóde zovšeobecnených odhadovacích rovníc (generalized estimating equations, GEE). Ako sme ukázali v predchádzajúcej kapitole, odhady regresných parametrov v zovšeobecnenom lineárnom modeli sú založené na princípe maximálnej vierohodnosti. Keďže metóda maximálnej vierohodnosti predpokladá, že podmienené rozdelenie pozorovaní je známe, bolo potrebné nájsť metódu, ktorá dokáže odhadnúť regresné koeficienty aj pre pozorovania, u ktorých nepoznáme ich podmienené rozdelenie. Preto metóda GEE nie je založená na maximálnej vierohodnosti, ale na princípe, ktorý kladie pre pozorovania slabšie predpoklady a dokáže pracovať s nesprávne špecifikovaným modelom.

Uvažujme teda K medzi sebou nezávislých náhodných vektorov $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_K$, kde každý z vektorov $\boldsymbol{Y}_i$, $i = 1, \ldots, K$ môžeme zapísať vo forme

$$\boldsymbol{Y}_i = (Y_{i1}, \ldots, Y_{in_i})^\top.$$

Náhodný vektor $\boldsymbol{Y}_i$ pozostáva z n_i závislých pozorovaní $\left(\sum_{i=1}^{K} n_i = N\right)$, ktoré náležia i-tej skupine. Naše dáta teda pozostávajú z K navzájom nezávislých skupín, ktorých pozorovania sú v jednotlivých skupinách medzi sebou závislé, ale sú nezávislé na pozorovaniach v ostatných skupinách. Takýmto dátam hovoríme *skupinovo závislé dáta*.

Ďalej označme $\boldsymbol{X}_{ij} = (X_{ij1}, \ldots, X_{ijp})^\top$ vektor vysvetľujúcich premenných o veľkosti p pre pozorovanie Y_{ij}. Rovnako, ako sme uvažovali v prípade zovšeobecneného lineárneho modelu, aj teraz chceme vyjadriť závislosť podmienenej strednej hodnoty Y_{ij} za podmienky $\boldsymbol{X}_{ij}$, označme ju $\mu_{ij} = \mathbb{E}[Y_{ij}|\boldsymbol{X}_{ij}]$, na vysvetľujúcich premenných $\boldsymbol{X}_{ij}$. Taktiež budeme predpokladať, že táto závislosť je tvaru $g(\mu_{ij}) = \eta_{ij} = \boldsymbol{X}_{ij}^\top \boldsymbol{\beta}$, kde g je rýdzo monotónna, dvakrát diferencovateľná linková funkcia a $\boldsymbol{\beta}$ je neznámy vektor skutočných regresných koeficientov. Z vyššie uvedeného vieme vyjadriť strednú hodnotu náhodného vektoru $\boldsymbol{Y}_i$ vo vektorovom zápise

$$\mathbb{E}\left[\boldsymbol{Y}_i|\boldsymbol{X}_{i1}, \ldots, \boldsymbol{X}_{in_i}\right] = \boldsymbol{\mu}_i = (\mu_{i1}, \ldots, \mu_{in_i})^\top,$$

kde hodnotu μ_{ij} môžeme vyjadriť pomocou inverznej funkcie k linkovej funkcii g^{-1} a to vzťahom $\mu_{ij} = g^{-1}(\eta_{ij}) = g^{-1}(\boldsymbol{X}_{ij}^\top \boldsymbol{\beta})$.

Kovariančnú maticu $var(\boldsymbol{Y}_i|\mathbb{X}_i)$ nebudeme nijako bližšie špecifikovať. Nebudeme klásť žiadne podmienky pre rozptyl a kovariancie medzi jednotlivými pozorovaniami v danej skupine. Namiesto skutočnej kovariančnej matice budeme uvažovať jej pracovnú verziu, ktorá sa môže od skutočnej kovariančnej matice líšiť. Táto pracovná verzia bude našim odhadom, ako by mohla skutočná kovariančná matica vyzerať.

Článok, v ktorom je prvýkrát predstavená metóda zovšeobecnených odhadovacých rovníc, je článok autorov Liang a Zeger (1986). V tomto článku je GEE

metóda odvodená pre longitudinálne data, ktoré patria medzi skupinovo závislé dáta. Ďalšie publikácie, na ktoré sa budeme odkazovať, sú Hardin a Hilbe (2003) a Ziegler (2011), ktoré sa taktiež venujú problematike GEE metódy. V tejto kapitole najskôr ukážeme, ako vyzerá tvar zovšeobecnenej odhadovacej rovnice. Ďalej predstavíme metódy, ktoré nás nakoniec privedú k samotnej metóde GEE predstavenej v článku Liang a Zeger (1986).

2.1 Tvar odhadovacej rovnice

V tejto časti predstavíme tvar zovšeobecnenej odhadovacej rovnice, ktorá pre odhad regresných koeficientov uvažuje taktiež aj koreláciu medzi jednotlivými pozorovaniami v skupine.

Na začiatok zadefinujeme matice, ktoré sa vyskytnú v tvare zovšeobecnenej odhadovacej rovnice. Označme

$$\mathbb{X}_i = \begin{pmatrix} \boldsymbol{X}_{i1}^\top \\ \vdots \\ \boldsymbol{X}_{in_i}^\top \end{pmatrix}$$

regresnú maticu náležiacu i-tej skupine pozorovaní. Ďalej označme

$$\mathbb{D}_i = \left(\frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}}\right) = \begin{pmatrix} g'(\mu_{i1}) & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & g'(\mu_{in_i}) \end{pmatrix}^{-1} \mathbb{X}_i.$$

maticu parciálnych derivácii vektoru $\boldsymbol{\mu}_i$ podľa zložiek vektoru neznámych koeficientov $\boldsymbol{\beta}$. Prvok na mieste (j,k) matice $\mathbb{D}_i$ dostaneme výpočtom

$$\frac{\partial \mu_{ij}}{\partial \beta_k} = \frac{\partial \mu_{ij}}{\partial \eta_{ij}} \frac{\partial \eta_{ij}}{\partial \beta_k} = \frac{1}{g'(\mu_{ij})} X_{ijk}.$$

Ďalej označme pracovnú kovariančnú maticu $\mathbb{V}_i$ náhodného vektoru $\boldsymbol{Y}_i$. Maticu $\mathbb{V}_i$ nazývame pracovnou kovariančnou maticou z toho dôvodu, že nepoznáme skutočnú kovariančnú maticu $var(\boldsymbol{Y}_i|\mathbb{X}_i) = \boldsymbol{\Sigma}_i$ a snažíme sa k nej priblížiť čo najviac, aj keď náš odhad nemusí byť správny.

Teraz ale prejdime k motivácii, ako sa dostaneme k tvaru zovšeobecnenej odhadovacej rovnice. Uvažujme, že združené rozdelenie vektoru $\boldsymbol{Y}_i$ je normálne so strednou hodnotou $\boldsymbol{\mu}_i$ a kovariančnou maticou $\boldsymbol{\Sigma}_i$, ktorá je známa. Hľadáme teda odhady regresných koeficientov $\boldsymbol{\beta}$ v klasickom lineárnom modeli. K odhadu $\boldsymbol{\beta}$ použijeme metódu zovšeobecnených minimálnych štvorcov (generalized least squares, GLS) viď Kariya (2004), kapitola 2. Táto metóda nájde odhad $\boldsymbol{\beta}$ minimalizáciou funkcie

$$\sum_{i=1}^{K} (\boldsymbol{Y}_i - \mathbb{X}_i\boldsymbol{\beta})^\top \boldsymbol{\Sigma}_i^{-1} (\boldsymbol{Y}_i - \mathbb{X}_i\boldsymbol{\beta}). \tag{2.1}$$

Derivovaním výrazu uvedeného v (2.1) podľa $\boldsymbol{\beta}$ vidíme, že ak minimum tejto funkcie existuje, tak rieši taktiež sústavu

$$\sum_{i=1}^{K} \mathbb{X}_i^\top \boldsymbol{\Sigma}_i^{-1} (\boldsymbol{Y}_i - \boldsymbol{\mu}_i) \stackrel{!}{=} \mathbf{0}_{p\times 1}, \tag{2.2}$$

kde $\boldsymbol{\mu}_i = \mathbb{X}_i\boldsymbol{\beta}$. Z výrazu (2.2) potom vieme explicitne vyjadriť tvar odhadu $\boldsymbol{\beta}$.

Obdobný postup môžeme použiť k nájdeniu odhadu $\boldsymbol{\beta}$ v zovšeobecnenej odhadovacej rovnici, ktorý je uvedený v Fitzmaurice a kol. (2004). Budeme derivovať výraz

$$\frac{1}{2} \cdot \sum_{i=1}^{K} (\boldsymbol{Y}_i - \boldsymbol{\mu}_i)^\top \mathbb{V}_i^{-1} (\boldsymbol{Y}_i - \boldsymbol{\mu}_i) \tag{2.3}$$

podľa $\boldsymbol{\beta}$, kde namiesto skutočnej kovariančnej matice $\boldsymbol{\Sigma}_i$ budeme uvažovať jej pracovnú verziu $\mathbb{V}_i$. Pre jednoduchosť budeme predpokladať, že matica $\mathbb{V}_i$ bude pevne daná a pre jednoduchosť nebude závislá na parametri $\boldsymbol{\beta}$. Deriváciou výrazu v (2.3) podľa $\boldsymbol{\beta}$ dostaneme tvar zovšeobecnenej odhadovacej rovnice, ktorý bol prvýkrát predstavený v článku Liang a Zeger (1986) ako

$$\boldsymbol{U}^K(\boldsymbol{\beta}) = \sum_{i=1}^{K} \mathbb{D}_i^\top \mathbb{V}_i^{-1} (\boldsymbol{Y}_i - \boldsymbol{\mu}_i) \stackrel{!}{=} \mathbf{0}_{p\times 1}. \tag{2.4}$$

Odhad $\widehat{\boldsymbol{\beta}}$ skutočného parametru $\boldsymbol{\beta}$ je riešením sústavy (2.4). Môžeme si všimnúť, že tvar skórovej funkcie v (2.4) je podobný skórovej funkcii v (1.9), ak by sme ju prepísali do vektorového tvaru.

V nadchádzajúcich častiach tejto práce sa zameriame na dve metódy, ktoré nás posunú k metóde GEE. Jedná sa o metódy pseudo maximálnej vierohodnosti a kvázi pseudo maximálnej vierohodnosti. Obe metódy sú uvedené v publikácii Ziegler (2011), odkiaľ budeme čerpať.

2.2 Metóda pseudo maximálnej vierohodnosti

V prvej kapitole sme ukázali, ako dokážeme odhadnúť regresné parametre $\boldsymbol{\beta}$ v zovšeobecnenom lineárnom modeli metódou maximálnej vierohodnosti. V tejto podkapitole naviažeme na túto metódu s tým, že oproti GLM nepoznáme skutočné podmienené rozdelenie $\boldsymbol{Y}_i$. Namiesto skutočného podmieneného rozdelenia budeme uvažovať nejaké pseudo rozdelenie, ktoré patrí do rodiny exponenciálych rozdelení. O takejto metóde hovoríme ako o metóde pseudo maximálnej vierohodnosti (pseudo maximum likelihood, PML). Prvýkrát bola predstavená v článku Gourieroux a kol. (1984).

Ako uvádza Ziegler (2011), budeme uvažovať K nezávislých n-rozmerných náhodných vektorov $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_K$ a $\mathbb{X}_i$ ako $n \times p$ regresnú maticu pre náhodný vektor $\boldsymbol{Y}_i$, $i = 1, \ldots, K$. Označme $f_T(\boldsymbol{Y}_i|\mathbb{X}_i)$ skutočné podmienené rozdelenie $\boldsymbol{Y}_i$ za podmienky $\mathbb{X}_i$. Keďže skutočné rozdelenie f_T nepoznáme, budeme predpokladať, že sa vektory pozorovaní $\boldsymbol{Y}_i$ budú riadiť nejakým rozdelením s hustotou $f_P(\boldsymbol{Y}_i|\mathbb{X}_i,\boldsymbol{\beta})$. Z dôvodu, že toto rozdelenie nie je skutočné, budeme ho nazývať pseudo rozdelením. Ďalej o tomto rozdelení predpokladajme, že patrí do rodiny exponenciálnych rozdelení s pevne daným parametrom $\boldsymbol{\Phi}_i$, ako je uvedené v časti 1.1.1.

Ďalej predpokladajme, že existuje vektor parametrov $\boldsymbol{\beta}$ taký, že podmienená stredná hodnota skutočného rozdelenia je rovná podmienenej strednej hodnote

pseudo rozdelenia, to je

$$\mathbb{E}_T[\boldsymbol{Y}_i|\mathbb{X}_i] = \mathbb{E}_P[\boldsymbol{Y}_i|\mathbb{X}_i,\boldsymbol{\beta}]. \tag{2.5}$$

Tento predpoklad je kľúčovým pre celú túto prácu. Pokiaľ nie je splnený, neplatia asymptotické výsledky uvedené v následujúcej časti.

Rovnosť v (2.5) nám hovorí, že model je správne špecifikovaný pre strednú hodnotu. Pre celú podkapitolu budeme predpokladať, že rovnosť v (2.5) platí. O skutočnej kovariančnej matici $\boldsymbol{\Sigma}_i$ budeme predpokladať, že existuje, ale nebudeme ju bližšie špecifikovať. Z predpokladu, že pseudo rozdelenie patrí do rodiny exponenciálnych rozdelení dostávame, že $var_P(\boldsymbol{Y}_i|\mathbb{X}_i) = \mathbb{V}_i$ existuje a je závislá na parametri $\boldsymbol{\beta}$.

Z predpokladu, že pseudo rozdelenie pochádza z rodiny exponenciálnych rozdelení, môžeme zapísať upravenú pseudo logaritmickú vierohodnostnú funkciu (pseudo loglikelihood function) ako

$$\ell_P^*(\boldsymbol{\beta}|\boldsymbol{Y}) = \sum_{i=1}^{K} \left(a(\boldsymbol{\mu}_i,\boldsymbol{\Phi}_i)^\top \boldsymbol{Y}_i + d(\boldsymbol{\mu}_i,\boldsymbol{\Phi}_i) \right), \tag{2.6}$$

kde je použitá parametrizácia vzhľadom ku strednej hodnote $\boldsymbol{\mu}_i$ tak, ako je uvedené v (1.3) a parameter $\boldsymbol{\mu}_i$ je závislý na neznámom parametri $\boldsymbol{\beta}$ a to vzťahom

$$\boldsymbol{\mu}_i = \boldsymbol{g}^{-1}(\mathbb{X}_i\boldsymbol{\beta}) = \begin{pmatrix} g^{-1}(\boldsymbol{X}_{i1}^\top\boldsymbol{\beta}) \\ \vdots \\ \vdots \\ g^{-1}(\boldsymbol{X}_{in_i}^\top\boldsymbol{\beta}) \end{pmatrix}.$$

Chceme nájsť taký odhad $\hat{\boldsymbol{\beta}}$ parametru $\boldsymbol{\beta}$, ktorý bude maximalizovať upravenú pseudo vierohodnostnú funkciu v (2.6). Derivovaním (2.6) podľa $\boldsymbol{\beta}$ dostávame

$$\begin{aligned}
\frac{\partial \ell_P^*(\boldsymbol{\beta}|\boldsymbol{Y})}{\partial \boldsymbol{\beta}} &= \sum_{i=1}^{K} \left(\frac{\partial a(\boldsymbol{\mu}_i,\boldsymbol{\Phi})^\top \boldsymbol{Y}_i}{\partial \boldsymbol{\beta}} + \frac{\partial d(\boldsymbol{\mu}_i,\boldsymbol{\Phi})}{\partial \boldsymbol{\beta}} \right) \\
&= \sum_{i=1}^{K} \left(\frac{\partial a(\boldsymbol{\mu}_i,\boldsymbol{\Phi})^\top \boldsymbol{Y}_i}{\partial \boldsymbol{\mu}_i} \frac{\partial \boldsymbol{\mu}_i^\top}{\partial \boldsymbol{\beta}} + \frac{\partial d(\boldsymbol{\mu}_i,\boldsymbol{\Phi})}{\partial \boldsymbol{\mu}_i} \frac{\partial \boldsymbol{\mu}_i^\top}{\partial \boldsymbol{\beta}} \right) \\
&= \sum_{i=1}^{K} \frac{\partial \boldsymbol{\mu}_i^\top}{\partial \boldsymbol{\beta}} \left(\frac{\partial a(\boldsymbol{\mu}_i,\boldsymbol{\Phi})^\top}{\partial \boldsymbol{\mu}_i} \boldsymbol{Y}_i + \frac{\partial d(\boldsymbol{\mu}_i,\boldsymbol{\Phi})}{\partial \boldsymbol{\mu}_i} \right) \\
&= \sum_{i=1}^{K} \frac{\partial \boldsymbol{\mu}_i^\top}{\partial \boldsymbol{\beta}} \frac{\partial a(\boldsymbol{\mu}_i,\boldsymbol{\Phi})^\top}{\partial \boldsymbol{\mu}_i} (\boldsymbol{Y}_i - \boldsymbol{\mu}_i) \\
&= \sum_{i=1}^{K} \mathbb{D}_i^\top \mathbb{V}_i^{-1} (\boldsymbol{Y}_i - \boldsymbol{\mu}_i),
\end{aligned}$$

kde sme v druhej rovnosti využili retiazkové pravidlo, vo štvrtej rovnosti sme využili tvrdenie 3 a v poslednej rovnosti sme využili tvrdenie 2 uvedené v prílohe A.1. Vidíme, že dostávame rovnaké vyjadrenie skórovej funkcie ako je uvedené v (2.4). Ukázali sme, že aj metódou pseudo maximálnej vierohodnosti sa vieme dostať k tvaru zovšeobecnenej odhadovacej rovnice. Odhad $\hat{\boldsymbol{\beta}}$ parametru $\boldsymbol{\beta}$, ktorý maximalizuje (2.6) nazveme pseudo maximálnym vierohodnostným odhadom (pseudo maximum likelihood estimator, PMLE).

2.2.1 Asymptotické vlastnosti odhadu PMLE

V tejto časti predstavíme asymptotické vlastnosti PML odhadu $\hat{\boldsymbol{\beta}}$. Na začiatok si zhrňme prepoklady, ktoré sme kládli na pozorovania:

- pseudo rozdelenie pochádza z rodiny exponenciálnych rozdelení,
- existuje $\boldsymbol{\beta}$ taký, že platí (2.5).

Ďalej v texte Ziegler (2011) autor predpokladá, že sú splnené podmienky regularity, ktoré sú uvedené v článku White (1982).

Za splnenie vyššie uvedených predpokladov má PMLE odhad $\hat{\boldsymbol{\beta}}$ parametru $\boldsymbol{\beta}$ podľa Ziegler (2011) (Theorem 5.2.) následujúce asymptotické vlastnosti:

1. Odhad $\hat{\boldsymbol{\beta}}$ konverguje skoro isto ku skutočnému parametru $\boldsymbol{\beta}$.
2. Odhad $\hat{\boldsymbol{\beta}}$ parametru $\boldsymbol{\beta}$ má asymptoticky normálne rozdelenie

$$\sqrt{K}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow[\mathrm{K} \to \infty]{\mathrm{d}} N_p\left(\mathbf{0}, \mathbb{B}^{-1}(\boldsymbol{\beta})\mathbb{M}(\boldsymbol{\beta})\mathbb{B}^{-1}(\boldsymbol{\beta})\right),$$

   kde

$$\mathbb{B}(\boldsymbol{\beta}) = \mathbb{E}\left[\mathbb{D}_i^\top \mathbb{V}_i^{-1}\mathbb{D}_i\right] \text{ a } \mathbb{M}(\boldsymbol{\beta}) = \mathbb{E}\left[\mathbb{D}_i^\top \mathbb{V}_i^{-1}\boldsymbol{\Sigma}_i\mathbb{V}_i^{-1}\mathbb{D}_i\right]$$

   a matice $\mathbb{D}_i$ a $\mathbb{V}_i$ sú matice definované v časti 2.1.

3. Matice $\mathbb{B}(\boldsymbol{\beta})$ a $\mathbb{M}(\boldsymbol{\beta})$ sa dajú konzistentne odhadnúť ako

$$\begin{aligned}\hat{\mathbb{B}}(\hat{\boldsymbol{\beta}}) &= \frac{1}{K}\sum_{i=1}^{K} \hat{\mathbb{D}}_i^\top \hat{\mathbb{V}}_i^{-1}\hat{\mathbb{D}}_i, \\ \hat{\mathbb{M}}(\hat{\boldsymbol{\beta}}) &= \frac{1}{K}\sum_{i=1}^{K} \hat{\mathbb{D}}_i^\top \hat{\mathbb{V}}_i^{-1}(\boldsymbol{Y}_i - \hat{\boldsymbol{\mu}}_i)(\boldsymbol{Y}_i - \hat{\boldsymbol{\mu}}_i)^\top \hat{\mathbb{V}}_i^{-1}\hat{\mathbb{D}}_i,\end{aligned} \tag{2.7}$$

   kde $\hat{\mathbb{D}}_i \equiv \mathbb{D}_i(\hat{\boldsymbol{\beta}})$, $\hat{\mathbb{V}}_i \equiv \mathbb{V}_i(\hat{\boldsymbol{\beta}})$ a $\hat{\boldsymbol{\mu}}_i \equiv \boldsymbol{\mu}_i(\hat{\boldsymbol{\beta}})$.

2.2.2 Príklady

Lineárna regresia s heteroskedasticitou

Uvažujme klasický lineárny model

$$Y_i = \boldsymbol{X}_i^\top \boldsymbol{\beta} + \varepsilon_i, \tag{2.8}$$

pre ktorý platí

$$\mathbb{E}[\varepsilon_i|\boldsymbol{X}_i] = 0 \qquad var\,[\varepsilon_i|\boldsymbol{X}_i] = \sigma^2(\boldsymbol{X}_i), \quad i = 1, \ldots, K,$$

kde $\boldsymbol{X}_i, \boldsymbol{\beta} \in \mathbb{R}^p$, $Y_1, \ldots, Y_K$ sú nezávislé a $\sigma^2(\boldsymbol{X}_i)$ je funkciou vysvetľujúcich premenných. Ďalej označme $\boldsymbol{Y} = (Y_1, \ldots, Y_K)^\top$ ako vektor pozorovaní, $\mathbb{X} = (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_K)^\top$ ako regresnú maticu a $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_K)^\top$ ako vektor chýb.

Nepoznáme skutočné rozdelenie Y_i, ale budeme predpokladať, že Y_i má normálne rozdelenie so strednou hodnotou $\mu_i = \boldsymbol{X}_i^\top \boldsymbol{\beta}$ a rozptylom $\sigma_i^2(\boldsymbol{X}_i) = \sigma^2$. Hustota podmieneného rozdelenia Y_i za podmienky $\boldsymbol{X}_i$ je tvaru

$$f_P(Y_i|\boldsymbol{X}_i,\boldsymbol{\beta}) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-(Y_i - \boldsymbol{X}_i^\top \boldsymbol{\beta})^2}{2\sigma^2}\right).$$

Odtiaľ môžeme písať upravenú pseudo logaritmickú vierohodnosť

$$\ell_P^*(\boldsymbol{\beta}|\boldsymbol{Y}) = -\sum_{i=1}^{K}(Y_i - \boldsymbol{X}_i^\top \boldsymbol{\beta})^2 - \frac{K}{2}\log(2\pi\sigma^2), \tag{2.9}$$

ktorú budeme maximalizovať vzhľadom k parametru $\boldsymbol{\beta}$. Deriváciou výrazu v (2.9) podľa $\boldsymbol{\beta}$ a položením výrazu nule dostávame

$$\frac{\partial \ell_P^*(\boldsymbol{\beta}|\boldsymbol{Y})}{\partial \beta} = \sum_{i=1}^{K} \boldsymbol{X}_i(Y_i - \boldsymbol{X}_i^\top \boldsymbol{\beta}) = \mathbb{X}^\top(\boldsymbol{Y} - \mathbb{X}\boldsymbol{\beta}) = \mathbf{0}_{p\times 1}.$$

PML odhad $\hat{\boldsymbol{\beta}}$ parametru $\boldsymbol{\beta}$ je rovný

$$\hat{\boldsymbol{\beta}} = (\mathbb{X}^\top\mathbb{X})^{-1}\mathbb{X}^\top\boldsymbol{Y},$$

čo je rovnaký tvar odhadu, ktorý dostávame metódou najmenších štvorcov pri odhade v klasickom lineárnom modeli za predpokladu homoskedasticity.

Z asymptotických vlastností PML odhadu uvedených v predchádzajúcej časti vidíme, že konzistentný odhad rozptylu PML odhadu $\hat{\boldsymbol{\beta}}$ je rovný

$$\widehat{var}_{PML}\,\hat{\boldsymbol{\beta}} = (\mathbb{X}^\top\mathbb{X})^{-1}\left(\mathbb{X}^\top(\boldsymbol{Y} - \mathbb{X}\hat{\boldsymbol{\beta}})(\boldsymbol{Y} - \mathbb{X}\hat{\boldsymbol{\beta}})^\top\mathbb{X}\right)(\mathbb{X}^\top\mathbb{X})^{-1}, \tag{2.10}$$

čo dostávame dosadením do (2.7) za

$$\mathbb{D} = \frac{\partial \boldsymbol{\mu}}{\partial \boldsymbol{\beta}} = \mathbb{X} \quad a \quad \mathbb{V}^{-1} = \frac{1}{\sigma^2}\mathbb{I}_K$$

a dosadením dostávame výraz ako je uvedený v (2.10). Tento odhad rozptylu je odlišný od odhadu rozptylu, ktorý je odvodený za predpokladu homoskedasticity

$$\widehat{var}_{OLS}\,\hat{\boldsymbol{\beta}} = \hat{\sigma}^2(\mathbb{X}^\top\mathbb{X})^{-1},$$

kde $\hat{\sigma}^2$ je odhad parametru σ^2. Môžeme si teda všimnúť, že aj napriek rovnakému tvaru odhadu je hodnota odhadnutého rozptylu pre PLM odhad $\hat{\boldsymbol{\beta}}$ rozdielna než pre OLS odhad za predpokladu homoskedasticity.

Odhadovacie rovnice s fixnou kovariančnou maticou

V tomto príklade ilustrujeme odhadovacie rovnice, ktorých pracovná kovariančná matica $\mathbb{V}_i$ bude fixná. To znamená, že budeme predpokladať, ako vyzerá skutočná kovariančná matica $\boldsymbol{\Sigma}_i$ a veríme, že naša pracovná kovariančná matica $\mathbb{V}_i$ je blízko k tejto reálne kovariančnej štruktúre.

Uvažujme K náhodných n-rozmerných vektorov $\boldsymbol{Y}_i$, $i = 1,\ldots,K$ a $\mathbb{X}_i$ maticu vysvetľujúcich premenných pre pozorovania v skupine i. Predpokladáme, že stredná hodnota je správne špecifikovaná a má tvar

$$\boldsymbol{\mu}_i = \mathbb{E}[\boldsymbol{Y}_i|\mathbb{X}_i,\boldsymbol{\beta}] = \boldsymbol{g}^{-1}(\mathbb{X}_i\boldsymbol{\beta}).$$

O skutočnej kovariančnej matici budeme predpokladať, že existuje, teda že existuje $var(\boldsymbol{Y}_i|\mathbb{X}_i)$, $i = 1,\ldots,K$. Žiadne ďalšie predpoklady pre skutočnú kovariančnú maticu klásť nebudeme.

Keďže nepoznáme skutočné rozdelenie náhodných vektorov $\boldsymbol{Y}_1,\ldots,\boldsymbol{Y}_K$, budeme predpokladať, že podmienené pseudo rozdelenie $\boldsymbol{Y}_i$ za podmienky $\mathbb{X}_i$ má normálne rozdelenie so strednou hodnotou $\boldsymbol{g}^{-1}(\mathbb{X}_i\boldsymbol{\beta})$ a s fixnou kovariančnou maticou $\mathbb{V}_i$ o veľkosti $n \times n$, teda

$$\boldsymbol{Y}_i|\mathbb{X}_i \sim N_n(\boldsymbol{g}^{-1}(\mathbb{X}_i\boldsymbol{\beta}),\mathbb{V}_i).$$

Môžeme písať pseudo logaritmickú vierohodnosť ako

$$\ell_P^*(\boldsymbol{\beta}|\boldsymbol{Y}) = -\frac{1}{2}\sum_{i=1}^{K}\left(\boldsymbol{Y}_i - \boldsymbol{g}^{-1}(\mathbb{X}_i\boldsymbol{\beta})\right)^\top \mathbb{V}_i^{-1}\left(\boldsymbol{Y}_i - \boldsymbol{g}^{-1}(\mathbb{X}_i\boldsymbol{\beta})\right). \tag{2.11}$$

Deriváciou pseudo logaritmickej vierohodnosti v (2.11) podľa $\boldsymbol{\beta}$ dostávame tvar odhadovacej rovnice

$$U^K(\boldsymbol{\beta}) = \sum_{i=1}^{K}\mathbb{D}_i^\top\mathbb{V}_i^{-1}\left(\boldsymbol{Y}_i - \boldsymbol{g}^{-1}(\mathbb{X}_i\boldsymbol{\beta})\right) \stackrel{!}{=} \mathbf{0}_{p\times 1}, \tag{2.12}$$

kde $\mathbb{D}_i = \frac{\partial\boldsymbol{\mu}_i}{\partial\boldsymbol{\beta}} = \frac{\partial\boldsymbol{g}^{-1}(\mathbb{X}_i\boldsymbol{\beta})}{\partial\boldsymbol{\beta}}$. Odhad $\hat{\boldsymbol{\beta}}$ parametru $\boldsymbol{\beta}$ je riešením sústavy (2.12).

Špecifickým prípadom je voľba jednotkovej matice $\mathbb{I}_n$ za pracovnú kovariančnú maticu $\mathbb{V}_i$. Takýmto odhadovacím rovniciam sa hovorí nezávislé odhadovacie rovnice s jednotkovou kovariančnou maticou. Využitie týchto odhadovacích rovníc nemusí byť v praxi prínosné, pretože reálne bude skutočná kovariančná matica rôzna od jednotkovej matice. Odhad $\hat{\boldsymbol{\beta}}$ parametru $\boldsymbol{\beta}$ založený na týchto rovniciach sa môže vylepšiť použitím kovariančnej matice, ktorá bude "bližšie"ku skutočnej korelačnej matici. Taktiež by sa mohla použiť kovariančná matica odhadnutá z dát. Tomuto prípadu sa bude venovať následujúca časť práce.

2.3 Metóda kvázi pseudo maximálnej vierohodnosti

V časti 2.2 sme predstavili PLM metódu, v ktorej sme nemuseli odhadovať žiaden parameter $\boldsymbol{\Phi}_i$, pretože sme uvažovali, že pre dané pseudo rozdelenie je tento parameter pevne daný. Tento prípad ale nie je veľmi častý, pretože nami pevne zvolená hodnota parametru $\boldsymbol{\Phi}_i$ nemusí byť blízka realite. To nás vedie k tomu, že je vhodné tento parameter odhadnúť.

Tak, ako v časti 2.2, aj teraz uvažujme K nezávislých n-rozmerných náhodných vektorov $\boldsymbol{Y}_1,\ldots,\boldsymbol{Y}_K$ a $\mathbb{X}_i$ ako $n \times p$ regresnú maticu pre náhodný vektor $\boldsymbol{Y}_i$, $i = 1,\ldots,K$. Opäť označme $f_T(\boldsymbol{Y}_i|\mathbb{X}_i)$ skutočné podmienené rozdelenie $\boldsymbol{Y}_i$ za podmienky $\mathbb{X}_i$. Predpokladáme, že sa pozorovania $\boldsymbol{Y}_i$ budú riadiť nejakým rozdelením s hustotou $f_Q(\boldsymbol{Y}_i|\mathbb{X}_i,\boldsymbol{\beta},\boldsymbol{\Phi}_i)$ patriaciou do rodiny exponenciálnych rozdelení. Stále budeme predpokladať, že stredná hodnota je správne špecifikovaná, to znamená, že existuje $\boldsymbol{\beta}$ taký, že platí

$$\mathbb{E}_T[\boldsymbol{Y}_i|\mathbb{X}_i] = \mathbb{E}_Q[\boldsymbol{Y}_i|\mathbb{X}_i,\boldsymbol{\beta},\boldsymbol{\Phi}_i]. \tag{2.13}$$

O skutočnej kovariančnej matici $\boldsymbol{\Sigma}_i$ opäť predpokladáme, že existuje, ale bližšie ju nešpecifikujeme. Namiesto skutočnej kovariančnej matice $\boldsymbol{\Sigma}_i$ budeme uvažovať pracovnú kovariančnú maticu $\mathbb{V}_i$, ktorú poznáme na základe pseudo rozdelenia. Tá je závislá nie len na parametri $\boldsymbol{\beta}$, ale taktiež na parametri $\boldsymbol{\alpha}$. Aký je význam parameteru $\boldsymbol{\alpha}$ a ako vyzerá jeho odhad $\hat{\boldsymbol{\alpha}}$ ukážeme v časti 2.4.1.

Aby sme mohli pristúpiť k odhadu parametru $\boldsymbol{\beta}$, potrebujeme najprv odhadnúť neznámy parameter $\boldsymbol{\Phi}_i$. Predpokladáme, že $\boldsymbol{\Phi}_i$ pre rozdelenia z exponenciálnej rodiny rozdelení môžeme podľa Ziegler (2011), časť 6.1, vyjadriť nejakou diferencovateľnou funkciou G, ktorá závisí na $\mathbb{X}_i$ a na neznámych parametroch $\boldsymbol{\beta}$ a $\boldsymbol{\alpha}$ ako

$$\boldsymbol{\Phi}_i = G(\mathbb{X}_i,\boldsymbol{\beta},\boldsymbol{\alpha}).$$

Hľadanie odhadu $\hat{\boldsymbol{\beta}}$ parametru $\boldsymbol{\beta}$ potom prebehne v následujúcich krokoch:

- nájdeme inicializačné konzistentné odhady $\tilde{\boldsymbol{\beta}}$, $\tilde{\boldsymbol{\alpha}}$ parametrov $\boldsymbol{\beta}$, $\boldsymbol{\alpha}$. Odhad $\tilde{\boldsymbol{\beta}}$ parametru $\boldsymbol{\beta}$ môžeme získať ako odhad GLM, ktorý je popísaný v časti 1.3. Na základe tohto odhadu potom vieme odhadnúť $\tilde{\boldsymbol{\alpha}}$ (ako odhadnúť $\tilde{\boldsymbol{\alpha}}$ v špecifických prípadoch ukážeme v následujúcej časti práce). Na základe odhadov $\tilde{\boldsymbol{\beta}}$ a $\tilde{\boldsymbol{\alpha}}$ dostávame odhad pre $\boldsymbol{\Phi}_i$ ako

  $$\tilde{\boldsymbol{\Phi}}_i = G(\mathbb{X}_i,\tilde{\boldsymbol{\beta}},\tilde{\boldsymbol{\alpha}}).$$

- dosadením odhadu $\tilde{\boldsymbol{\Phi}}_i$ budeme hľadať parameter $\boldsymbol{\beta}$ taký, ktorý maximalizuje výraz

  $$\ell_P^*(\boldsymbol{\beta},\tilde{\boldsymbol{\alpha}},|\boldsymbol{Y}) = \sum_{i=1}^{K}\left(a(\boldsymbol{\mu}_i,\tilde{\boldsymbol{\Phi}}_i)^\top\boldsymbol{Y}_i + d(\boldsymbol{\mu}_i,\tilde{\boldsymbol{\Phi}}_i)\right). \tag{2.14}$$

  K tomuto výrazu sa dostaneme z predpokladu, že pseudo rozdelenie pochádza z rodiny exponenciálnych rozdelení.

Výrazu v (2.14) hovoríme kvazi pseudo logaritmická vierohodnosť. Je určitým rozšírením pseudo logaritmickej vierohodnosti s tým rozdielom, že parameter $\boldsymbol{\Phi_i}$ je neznámym parametrom a je potreba ho odhadnúť.

Ukážeme, že aj týmto prístupom sa dokážeme dostať k rovnakému tvaru zovšeobecnenej odhadovacej rovnice ako v (2.4). Počítajme

$$\frac{\partial\ell_P^*(\boldsymbol{\beta},\tilde{\boldsymbol{\alpha}}|\boldsymbol{Y})}{\partial\boldsymbol{\beta}} = \sum_{i=1}^{K}\left(\frac{\partial a(\boldsymbol{\mu}_i,\tilde{\boldsymbol{\Phi}}_i)^\top\boldsymbol{Y}_i}{\partial\boldsymbol{\beta}} + \frac{\partial d(\boldsymbol{\mu}_i,\tilde{\boldsymbol{\Phi}}_i)}{\partial\boldsymbol{\beta}}\right)$$

$$
\begin{aligned}
&= \sum_{i=1}^{K} \left(\frac{\partial a(\boldsymbol{\mu}_i, \tilde{\boldsymbol{\Phi}}_i)^\top \boldsymbol{Y}_i}{\partial \boldsymbol{\mu}_i} \frac{\partial \boldsymbol{\mu}_i^\top}{\partial \boldsymbol{\beta}} + \frac{\partial d(\boldsymbol{\mu}_i, \tilde{\boldsymbol{\Phi}}_i)}{\partial \boldsymbol{\mu}_i} \frac{\partial \boldsymbol{\mu}_i^\top}{\partial \boldsymbol{\beta}} \right) \\
&= \sum_{i=1}^{K} \frac{\partial \boldsymbol{\mu}_i^\top}{\partial \boldsymbol{\beta}} \left(\frac{\partial a(\boldsymbol{\mu}_i, \tilde{\boldsymbol{\Phi}}_i)^\top}{\partial \boldsymbol{\mu}_i} \boldsymbol{Y}_i + \frac{\partial d(\boldsymbol{\mu}_i, \tilde{\boldsymbol{\Phi}}_i)}{\partial \boldsymbol{\mu}_i} \right) \\
&= \sum_{i=1}^{K} \frac{\partial \boldsymbol{\mu}_i^\top}{\partial \boldsymbol{\beta}} \frac{\partial a(\boldsymbol{\mu}_i, \tilde{\boldsymbol{\Phi}}_i)^\top}{\partial \boldsymbol{\mu}_i} \left(\boldsymbol{Y}_i - \boldsymbol{\mu}_i \right) \\
&= \sum_{i=1}^{K} \mathbb{D}_i^\top \tilde{\mathbb{V}}_i^{-1} \left(\boldsymbol{Y}_i - \boldsymbol{\mu}_i \right),
\end{aligned}
$$

kde $\tilde{\mathbb{V}}_i$ je pracovná kovariančná matica, do ktorej sme dosadili odhad $\tilde{\boldsymbol{\Phi}}_i$. Pri úprave v druhom kroku sme použili retiazkové pravidlo a v štvrtom kroku sme aplikovali tvrdenie 3 a v poslednej rovnosti sme využili tvrdenie 2, ktoré sú uvedené v prílohe A.1.

Vidíme, že dostávame tvar zovšeobecnenej odhadovacej rovnice ako v (2.4) s tým rozdielom, že matica $\tilde{\mathbb{V}}_i$ je závislá na odhadnutých parametroch $\tilde{\boldsymbol{\alpha}}$ a $\tilde{\boldsymbol{\Phi}}_i$, kde $\boldsymbol{\Phi}_i$ je funkciou $\tilde{\boldsymbol{\alpha}}$ a $\tilde{\boldsymbol{\beta}}$. Odhad $\hat{\boldsymbol{\beta}}$ parametru $\boldsymbol{\beta}$ maximalizuje (2.14) a je riešením zovšeobecnenej odhadovacej rovnice. Nazývame ho kvazi pseudo maximálne vierohodný odhad (quasi pseudo maximum likelihood estimator, QPMLE).

2.3.1 Asymptotické vlastnosti odhadu QPMLE

V tejto časti uvedieme asymptotické vlastnosti QPML odhadu. K tomu, aby nižšie uvedné asymptotické výsledky platili, musia byť splnené následovné predpoklady:

- je splnená podmienka uvedená v (2.13),
- podmienené rozdelenie $f_Q(\boldsymbol{Y}_i|\mathbb{X}_i,\boldsymbol{\beta},\boldsymbol{\Phi}_i)$ pochádza z rodiny exponenciálnych rozdelení,
- Ziegler (2011) uvádza, že musia byť splnené podmienky regularity, ktoré sú uvedné v článku Gourieroux a kol. (1984).
- odhady $\tilde{\boldsymbol{\alpha}}$ a $\tilde{\boldsymbol{\Phi}}_i$ sú $K^{\frac{1}{2}}$-konzistentné odhady parametrov $\boldsymbol{\alpha}$ a $\boldsymbol{\Phi}_i$ (definícia $K^{\frac{1}{2}}$- konzistencie je uvedená v prílohe A.2).

Za splnenia predpokladov uvedených vyššie platia podľa Ziegler (2011) (Theorem 6.2) následujúce asymptotické vlastnosti QPML odhadu $\hat{\boldsymbol{\beta}}$ parametru $\boldsymbol{\beta}$:

1. Odhad $\hat{\boldsymbol{\beta}}$ konverguje skoro iste ku skutočnému parametru $\boldsymbol{\beta}$.
2. Odhad $\hat{\boldsymbol{\beta}}$ parametru $\boldsymbol{\beta}$ má asymptoticky normálne rozdelenie

$$
\sqrt{K}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow[\mathrm{K} \to \infty]{\mathrm{d}} N_p \left(\mathbf{0}, \mathbb{B}^{-1}(\boldsymbol{\beta},\boldsymbol{\alpha}) \mathbb{M}(\boldsymbol{\beta},\boldsymbol{\alpha}) \mathbb{B}^{-1}(\boldsymbol{\beta},\boldsymbol{\alpha}) \right),
$$

kde

$$
\mathbb{B}(\boldsymbol{\beta},\boldsymbol{\alpha}) = \mathbb{E}\left[\mathbb{D}_i^\top \mathbb{V}_i^{-1} \mathbb{D}_i \right] \text{ a } \mathbb{M}(\boldsymbol{\beta},\boldsymbol{\alpha}) = \mathbb{E}\left[\mathbb{D}_i^\top \mathbb{V}_i^{-1} \boldsymbol{\Sigma}_i \mathbb{V}_i^{-1} \mathbb{D}_i \right]
$$

a matice $\mathbb{D}_i$ a $\mathbb{V}_i$ sú matice definované v časti 2.1.

3. Matice $\mathbb{B}(\boldsymbol{\beta},\boldsymbol{\alpha})$ a $\mathbb{M}(\boldsymbol{\beta},\boldsymbol{\alpha})$ sa dajú konzistentne odhadnúť ako

$$\hat{\mathbb{B}}(\hat{\boldsymbol{\beta}},\tilde{\boldsymbol{\alpha}}) = \frac{1}{K}\sum_{i=1}^{K}\hat{\mathbb{D}}_i^\top\hat{\mathbb{V}}_i^{-1}\hat{\mathbb{D}}_i,$$
$$\hat{\mathbb{M}}(\hat{\boldsymbol{\beta}},\tilde{\boldsymbol{\alpha}}) = \frac{1}{K}\sum_{i=1}^{K}\hat{\mathbb{D}}_i^\top\hat{\mathbb{V}}_i^{-1}(\boldsymbol{Y}_i-\hat{\boldsymbol{\mu}}_i)(\boldsymbol{Y}_i-\hat{\boldsymbol{\mu}}_i)^\top\hat{\mathbb{V}}_i^{-1}\hat{\mathbb{D}}_i,$$

kde $\hat{\mathbb{D}}_i \equiv \mathbb{D}_i(\hat{\boldsymbol{\beta}},\tilde{\boldsymbol{\alpha}})$, $\hat{\mathbb{V}}_i \equiv \mathbb{V}_i(\hat{\boldsymbol{\beta}},\tilde{\boldsymbol{\alpha}})$ a $\hat{\boldsymbol{\mu}}_i \equiv \boldsymbol{\mu}_i(\hat{\boldsymbol{\beta}})$.

2.3.2 Príklady

Odhadovacie rovnice s odhadnutou kovariančnou maticou

V časti 2.2.2 sme ukázali, ako vyzerá odhadovacia rovnica v prípade, že uvažujeme fixnú pracovnú kovariančnú maticu $\mathbb{V}_i$. V skutočnosti ale nevieme, či je náš odhad pracovnej kovariančnej matice správny. Môžeme mať ale informáciu, ako by mohla vyzerať štruktúra skutočnej kovariančnej matice $\boldsymbol{\Sigma}_i$. Na základe tejto informácie môžeme odhadnúť pracovnú kovariančnú maticu $\mathbb{V}_i$.

Ako v časti 2.2.2, opäť uvažujme K náhodných n-rozmerných vektorov $\boldsymbol{Y}_i$, $i = 1,\ldots,K$. Ďalej označme $\mathbb{X}_i$ maticu vysvetľujúcich premenných pre pozorovania v skupine i. Predpokladáme, že stredná hodnota je správne špecifikovaná a má tvar

$$\boldsymbol{\mu}_i = \mathbb{E}[\boldsymbol{Y}_i|\mathbb{X}_i,\boldsymbol{\beta}] = \boldsymbol{g}^{-1}(\mathbb{X}_i\boldsymbol{\beta}).$$

O skutočnej kovariančnej matici budeme predpokladať, že existuje, teda že existuje $var(\boldsymbol{Y}_i|\mathbb{X}_i)$, $i = 1,\ldots,K$. Ďalej budeme predpokladať, že máme informáciu, ktorá nám hovorí o tom, ako by mohla vyzerať štruktúra skutočnej kovariančnej matice, napríklad

$$var\,(Y_{ij}) = \sigma_1^2 \quad a \quad cov\,(Y_{ij},Y_{ik}) = \sigma_2^2,$$

kde $i = 1,\ldots,K$, $j,k = 1,\ldots,n$, $j \neq k$. Najskôr nájdeme inicializačný odhad $\tilde{\boldsymbol{\beta}}$, na základe ktorého nájdeme vhodné odhady $\hat{\sigma}_1^2$ a $\hat{\sigma}_2^2$. Z týchto odhadov a informácie o štruktúre skutočnej kovariančnej matice dostávame odhadnutú pracovnú kovariančnú maticu $\hat{\mathbb{V}}_i$.

Keďže nepoznáme skutočné podmienené rozdelenie $\boldsymbol{Y}_i$, budeme uvažovať, že podmienené rozdelenie $\boldsymbol{Y}_i$ za podmienky $\mathbb{X}_i$ sa riadi normálnym rozdelením so strednou hodnotou $\boldsymbol{\mu}_i = \boldsymbol{g}^{-1}(\mathbb{X}_i\boldsymbol{\beta})$ a kovariančnou maticou $\hat{\mathbb{V}}_i$, teda

$$\boldsymbol{Y}_i|\mathbb{X}_i \sim N_n(\boldsymbol{\mu}_i,\hat{\mathbb{V}}_i)$$

Ďalej môžeme pokračovať ako v príklade v časti 2.2.2. Môžeme písať kvazi pseudo logaritmickú vierohodnosť ako

$$\ell_P^*(\boldsymbol{\beta}|\boldsymbol{Y}) = -\frac{1}{2}\sum_{i=1}^{K}\left(\boldsymbol{Y}_i-\boldsymbol{g}^{-1}(\mathbb{X}_i\boldsymbol{\beta})\right)^\top\hat{\mathbb{V}}_i^{-1}\left(\boldsymbol{Y}_i-\boldsymbol{g}^{-1}(\mathbb{X}_i\boldsymbol{\beta})\right). \tag{2.15}$$

Deriváciou výrazu v (2.15) podľa $\boldsymbol{\beta}$ dostávame tvar odhadovacej rovnice

$$U^K(\boldsymbol{\beta}) = \sum_{i=1}^{K} \mathbb{D}_i^\top \hat{\mathbb{V}}_i^{-1} \left(\boldsymbol{Y}_i - \boldsymbol{g}^{-1}(\mathbb{X}_i \boldsymbol{\beta}) \right) \stackrel{!}{=} \mathbf{0}_{p\times 1}. \tag{2.16}$$

Odhad $\hat{\boldsymbol{\beta}}$ parametru $\boldsymbol{\beta}$ je riešením sústavy (2.16).

Špeciálnym prípadom metódy kvazi pseudo maximálnej vierohodnosti je metóda GEE uvedená v Liang a Zeger (1986). Špeciálnym prípadom je z toho dôvodu, že pracovnú kovariančnú maticu uvažujeme ako

$$\mathbb{V}_i = \mathbb{A}_i^{\frac{1}{2}}(\boldsymbol{\beta}) \mathbb{R}_i(\boldsymbol{\alpha}) \mathbb{A}_i^{\frac{1}{2}}(\boldsymbol{\beta}),$$

kde $\mathbb{A}_i$ je diagonálna matica závislá na parametri $\boldsymbol{\beta}$ a matica $\mathbb{R}_i$ je pracovná korelačná matica, ktorá je závislá len na parametri $\boldsymbol{\alpha}$. Túto metódu predstavíme v nasledujúcej časti.

2.4 Metóda GEE

V predchádzajúcej časti sme ukázali, ako nájsť QPMLE odhad $\hat{\boldsymbol{\beta}}$ parametru $\boldsymbol{\beta}$ a aké sú jeho asymptotické vlastnosti v prípade, že je pracovná kovariančná matica $\mathbb{V}_i$ závislá na dodatočných parametroch $\boldsymbol{\alpha}$ a $\boldsymbol{\Phi}_i$. Metóda GEE predstavená v článku autorov Liang a Zeger (1986) predpokladá, že medzi pozorovaniami v skupine je určitá korelácia. V tejto časti sa zameráme práve na túto metódu. Ďalej sa v nej zameriame na závislosť pracovnej kovariančnej matice $\mathbb{V}_i$ na parametri $\boldsymbol{\alpha}$ prostredníctvom pracovnej korelačnej matice, ktorú budeme značiť $\mathbb{R}_i(\boldsymbol{\alpha})$.

Liang a Zeger (1986) vo svojom článku využívajú štruktúru GLM modelu a pracovnej korelačnej matice. Využívajú vzťahy

$$\mathbb{E}[Y_{ij}|\mathbb{X}_i] = g(\boldsymbol{X}_{ij}^\top \boldsymbol{\beta}) = \mu_{ij} \quad \text{a} \quad var\,[Y_{ij}|\mathbb{X}_i] = \varphi V(\mu_{ij}) = a_{ij},$$

kde $V(\mu_{ij})$ je hodnota variančnej funkcie v bode μ_{ij}, $i = 1, \ldots, K$, $j = 1, \ldots, n_i$ a ďalej

$$\mathbb{V}_i = \mathbb{A}_i^{\frac{1}{2}} \mathbb{R}_i(\boldsymbol{\alpha}) \mathbb{A}_i^{\frac{1}{2}},$$

kde $\mathbb{A}_i = \mathrm{diag}(a_{ij})$ je diagonálna matica s prvkami a_{ij} na diagonále, a $\mathbb{R}_i(\boldsymbol{\alpha})$ je pracovná korelačná matica závislá na parametri $\boldsymbol{\alpha}$. V praxi je táto pracovná korelačná matica volená jednotne pre všetky skupiny.

Ako sme uvideli v časti 2.1, zovšeobecnené odhadovacie rovnice sú tvaru ako v (2.4). Aby sme mohli nájsť odhad $\hat{\boldsymbol{\beta}}$ parametru $\boldsymbol{\beta}$ metódou GEE podľa Liang a Zeger (1986), potrebujeme nájsť taktiež odhad parametrov $\boldsymbol{\alpha}$ a φ. O odhadoch týchto parametrov predpokladajú, že sú $K^{\frac{1}{2}}$- konzistentné. Dosadením do odhadovacej rovnice v (2.4) dostávame

$$\sum_{i=1}^{K} \mathbb{D}_i^\top \tilde{\mathbb{V}}_i^{-1} (\boldsymbol{Y}_i - \boldsymbol{\mu}_i) = \mathbf{0}_{p\times 1},$$

kde $\tilde{\mathbb{V}}_i$ je závislá na už odhadnutých parametroch $\tilde{\boldsymbol{\alpha}}$ a $\hat{\varphi}$. Odhad $\hat{\boldsymbol{\beta}}$ parametru $\boldsymbol{\beta}$ je potom riešením tejto sústavy. Môžeme si všimnúť, že sa jedná o určitú modifikáciu metódy kvazi pseudo maximálnej vierohodnosti, kde naviac ešte predpokladáme $K^{\frac{1}{2}}$-konzistenciu odhadov parametrov $\boldsymbol{\alpha}$ a φ.

Fitzmaurice a kol. (2004) vo svojej práci uvádzajú asymptotické vlastnosti odhadu $\hat{\boldsymbol{\beta}}$ parametru $\boldsymbol{\beta}$ založeného na metóde GEE podľa Liang a Zeger (1986):

1. Odhad $\hat{\boldsymbol{\beta}}$ konverguje skoro iste ku skutočnému parametru $\boldsymbol{\beta}$.

2. Odhad $\hat{\boldsymbol{\beta}}$ parametru $\boldsymbol{\beta}$ má asymptoticky normálne rozdelenie

$$\sqrt{K}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow[\mathrm{K} \to \infty]{\mathrm{d}} N_p\left(\mathbf{0}, \mathbb{B}^{-1}(\boldsymbol{\beta},\boldsymbol{\alpha})\mathbb{M}(\boldsymbol{\beta},\boldsymbol{\alpha})\mathbb{B}^{-1}(\boldsymbol{\beta},\boldsymbol{\alpha})\right), \tag{2.17}$$

   kde

$$\mathbb{B}(\boldsymbol{\beta},\boldsymbol{\alpha}) = \lim_{K\to\infty} \frac{1}{K} \sum_{i=1}^{K} \mathbb{D}_i^\top \mathbb{V}_i^{-1} \mathbb{D}_i \text{ a } \mathbb{M}(\boldsymbol{\beta},\boldsymbol{\alpha}) = \lim_{K\to\infty} \frac{1}{K} \sum_{i=1}^{K} \mathbb{D}_i^\top \mathbb{V}_i^{-1} \boldsymbol{\Sigma}_i \mathbb{V}_i^{-1} \mathbb{D}_i$$

   a matice $\mathbb{D}_i$ a $\mathbb{V}_i$ sú matice definované v časti 2.1.

3. Matice $\mathbb{B}(\boldsymbol{\beta})$ a $\mathbb{M}(\boldsymbol{\beta})$ sa dajú konzistentne odhadnúť ako

$$\hat{\mathbb{B}}(\hat{\boldsymbol{\beta}},\tilde{\boldsymbol{\alpha}}) = \frac{1}{K} \sum_{i=1}^{K} \hat{\mathbb{D}}_i^\top \hat{\mathbb{V}}_i^{-1} \hat{\mathbb{D}}_i,$$
$$\hat{\mathbb{M}}(\hat{\boldsymbol{\beta}},\tilde{\boldsymbol{\alpha}}) = \frac{1}{K} \sum_{i=1}^{K} \hat{\mathbb{D}}_i^\top \hat{\mathbb{V}}_i^{-1} (\boldsymbol{Y}_i - \hat{\boldsymbol{\mu}}_i)(\boldsymbol{Y}_i - \hat{\boldsymbol{\mu}}_i)^\top \hat{\mathbb{V}}_i^{-1} \hat{\mathbb{D}}_i,$$

   kde $\hat{\mathbb{D}}_i \equiv \mathbb{D}_i(\hat{\boldsymbol{\beta}},\tilde{\boldsymbol{\alpha}})$, $\hat{\mathbb{V}}_i \equiv \mathbb{V}_i(\hat{\boldsymbol{\beta}},\tilde{\boldsymbol{\alpha}})$ a $\hat{\boldsymbol{\mu}}_i \equiv \boldsymbol{\mu}_i(\hat{\boldsymbol{\beta}})$.

Odhad asymptotického rozptylu v (2.17) je známy ako empirický "sendvičový" odhad. Podľa Fitzmaurice a kol. (2004) (kapitola 11, strana 303) nám udáva konzistentné odhady asymptotického rozptylu odhadu $\widehat{\boldsymbol{\beta}}$, aj keď predpokladaná korelačná štruktúra nie je presne špecifikovaná.

Pri tomto zistení sa nám naskytá otázka, či je naozaj potrebné odhadovať skutočnú korelačnú štruktúru, keď získavame konzistentný odhad rozptylu odhadu $\hat{\boldsymbol{\beta}}$ aj bez správnej špecifikácie korelačnej štruktúry. Na túto otázku máme odpoveď - čím presnejšie sa nám podarí odhadnúť skutočnú korelačnú štruktúru, tým lepší odhad parametru $\boldsymbol{\beta}$ dostaneme. To znamená, že rozptyl ohadu $\hat{\boldsymbol{\beta}}$ parametru $\boldsymbol{\beta}$ bude menší pri presnejšie odhadnutej korelačnej štruktúre. Odtiaľ môžeme usúdiť, že čím lepší odhad $\hat{\boldsymbol{\beta}}$ parametru $\boldsymbol{\beta}$ chceme, tým viac sa musíme priblížiť ku skutočnej korelačnej štruktúre.

Výpočet odhadu $\widehat{\boldsymbol{\beta}}$ je numericky veľmi náročný, keďže neodhadujeme len parameter $\boldsymbol{\beta}$, ale aj parametre $\boldsymbol{\alpha}$ a φ, na ktorých je závislá matica $\mathbb{V}_i$. Odhad $\hat{\boldsymbol{\beta}}$ nájdeme algoritmom, ktorý postupne iteruje odhady $\widehat{\boldsymbol{\beta}}_k$, až kým nedôjde ku konvergencii. Tento algoritmus je založený na modifikácii Fisherovej skórovej metódy. Algoritmus popíšeme v pár stručných krokoch.

Iteratívny algoritmus k nájdeniu $\widehat{\boldsymbol{\beta}}$

1. Na začiatok špecifikujeme model, v ktorom budeme odhadovať regresné koeficienty $\boldsymbol{\beta}$. To znamená špecifikáciu strednej hodnoty $\boldsymbol{\mu}_i$, linkovej funkcie a štruktúry pracovnej korelačnej matice.

2. Nájdeme inicializačný parameter $\tilde{\boldsymbol{\beta}}$ ako odhad pomocou GLM modelu ako je uvedené v 1.3. Položme $\hat{\boldsymbol{\beta}}^{(0)} = \tilde{\boldsymbol{\beta}}$.

3. Na základe $\hat{\boldsymbol{\beta}}^{(k)}$ spočítame Personove reziduá

$$\hat{e}_{ij} = \frac{Y_{ij} - \widehat{\mu}_{ij}}{\sqrt{V(\widehat{\mu}_{ij})}},$$

   kde $\widehat{\mu_{ij}} = g(\boldsymbol{X}_{ij}^{\top}\hat{\boldsymbol{\beta}}^{(k)})$. Z Pearsonových reziduí spočítame odhad $\tilde{\boldsymbol{\alpha}}$ a odhad $\hat{\varphi}$. Odhad $\tilde{\boldsymbol{\alpha}}$ je pre rôzny výber korelačnej štruktúry iný. Tvar odhadov $\tilde{\boldsymbol{\alpha}}$ pre vybrané korelačné štruktúry uvedieme v následujúcej časti tejto práce, rovnako ako aj tvar odhadu $\hat{\varphi}$.

4. Spočítame odhad matice $\mathbb{V}_i$ ako

$$\tilde{\mathbb{V}}_i(\hat{\boldsymbol{\beta}}^{(k)},\tilde{\boldsymbol{\alpha}}) = \tilde{\mathbb{A}}_i^{\frac{1}{2}}\mathbb{R}(\tilde{\boldsymbol{\alpha}})\tilde{\mathbb{A}}_i^{\frac{1}{2}},$$

   kde $(\tilde{\mathbb{A}}_i)_{[j,j]} = \hat{\varphi}V(\hat{\mu}_{ij})$ a $V(\hat{\mu}_{ij})$ je hodnota rozptylovej funkcie v bode $\widehat{\mu}_{ij} = \boldsymbol{X}_{ij}^{\top}\hat{\boldsymbol{\beta}}^{(k)}$ a $\mathbb{R}_i(\tilde{\boldsymbol{\alpha}})$ je vhodne zvolená korelačná štruktúra.

5. Zaktualizujeme hodnotu odhadu $\widehat{\boldsymbol{\beta}}^{(k)}$ pomocou iteračnej formuly

$$\widehat{\boldsymbol{\beta}}^{(k+1)} = \widehat{\boldsymbol{\beta}}^{(k)} + \left[\sum_{i=1}^{K}\tilde{\mathbb{D}}_i^{\top}\tilde{\mathbb{V}}_i^{-1}\tilde{\mathbb{D}}_i\right]^{-1}\sum_{i=1}^{K}\tilde{\mathbb{D}}_i^{\top}\tilde{\mathbb{V}}_i^{-1}(\boldsymbol{Y}_i - \widehat{\boldsymbol{\mu}}_i),$$

   kde $\tilde{\mathbb{D}}_i \equiv \mathbb{D}_i(\widehat{\boldsymbol{\beta}}^{(k)})$ je odhad matice $\mathbb{D}_i$ po dosadení odhadnutých regresných koeficientov $\widehat{\boldsymbol{\beta}}^{(k)}$.

6. Zaktualizujeme odhad $\tilde{\boldsymbol{\alpha}}$ na základe aktuálnej hodnoty odhadu $\hat{\boldsymbol{\beta}}^{(k+1)}$ (ako vyzerá odhad $\tilde{\boldsymbol{\alpha}}$ je uvedené v následujúcej časti práce).

7. Opakujeme kroky 3 až 6, pokiaľ $\|\widehat{\boldsymbol{\beta}}^{(k+1)} - \widehat{\boldsymbol{\beta}}^{(k)}\| < \delta$, kde δ je predom zvolená konštanta pre konvergenciu.

Ako sme uviedli v bode 3 iteratívneho algoritmu, v následujúcej časti práce predstavíme odhady $\tilde{\boldsymbol{\alpha}}$ parametru $\boldsymbol{\alpha}$ pre najpoužívanejšie korelačné štruktúry.

2.4.1 Parametrizácia pracovných korelačných matíc a odhad parametru α

Ako sme už v tejto práci spomínali, nepoznáme skutočnú koreláciu medzi jednotlivými pozorovaniami v skupine. Preto zavádzame pracovnú korelačnú maticu, ktorú potom môžeme použiť k odhadu skutočnej korelácie.

Pred tým, než sa dostaneme k predstaveniu rôznych korelačných štruktúr, pripomeňme si odhadnuté Pearsonove reziduá, ktoré sú tvaru

$$\hat{e}_{ij} = \frac{Y_{ij} - \widehat{\mu}_{ij}}{\sqrt{V(\widehat{\mu}_{ij})}},$$

kde odhady $\widehat{\mu}_{ij}$ získame z aktuálneho odhadu $\hat{\boldsymbol{\beta}}^{(k)}$ z iteratívneho algoritmu popísaného vyššie.

Pokiaľ má model správne špecifikovanú strednú hodnotu, potom pre odhadnuté Pearsonove reziduá dostávame

$$\mathbb{E}\hat{e}_{ij} \approx 0, \quad \operatorname{var} \hat{e}_{ij} \approx \varphi, \quad \mathbb{E}\hat{e}_{ij}\hat{e}_{ik} \approx \varphi\alpha_{jk},$$

kde α_{jk} je skutočná korelácia medzi pozorovaniami Y_{ij} a Y_{ik}. Vyššie uvedené platí vďaka asymptotickým vlastnostiam odhadu $\hat{\mu}_{ij}$, keďže predpokladáme, že $\hat{\mu}_{ij}$ je konzistentným odhadom μ_{ij}.

Teraz prejdeme k predstaveniu niektorých najpoužívanejších štruktúr pracovných korelačných matíc a odhadov príslušných parametrov α k daným štruktúram:

Nezávislá korelačná štruktúra

Táto korelačná štruktúra je najjednoduchšou z uvádzaných korelačných štruktúr. Uvažuje, že pozorovania sú medzi sebou v skupine nezávislé. Matica $\mathbb{R}_i(\boldsymbol{\alpha})$ je tvaru

$$\mathbb{R}_i(\boldsymbol{\alpha}) = \begin{pmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 1 \end{pmatrix}.$$

V tomto prípade teda nie je potrebné odhadovať žiaden ďalší parameter $\boldsymbol{\alpha}$.

Spoločná korelačná štruktúra

V tejto štruktúre predpokladáme, že medzi všetkými pozorovaniami v skupine je rovnaká korelácia. To znamená, že všetky prvky matice $\mathbb{R}_i(\boldsymbol{\alpha})$ okrem hlavnej diagonály sú rovnaké. Túto štruktúru môžeme zapísať ako

$$(\mathbb{R}_i)_{[k,l]} = \begin{cases} 1, & k = l \\ \alpha, & k \neq l, \end{cases} \tag{2.18}$$

kde $\alpha \in (0,1)$. Matica $\mathbb{R}_i(\boldsymbol{\alpha})$ bude tvaru

$$\mathbb{R}_i(\boldsymbol{\alpha}) = \begin{pmatrix} 1 & \alpha & \dots & \alpha \\ \alpha & 1 & \dots & \alpha \\ \vdots & \vdots & \ddots & \vdots \\ \alpha & \alpha & \dots & 1 \end{pmatrix}. \tag{2.19}$$

Spoločnú korelačnú štruktúru môžeme použiť najmä pri dátach, ktoré nie sú časovo závislé a nezáleží na poradí pozorovaní.

V štruktúre (2.18) odhadujeme konštantu α a odhad $\tilde{\alpha}$ je tvaru

$$\tilde{\alpha} = \frac{1}{\hat{\varphi}} \frac{\sum_{i=1}^{K} \sum_{j<k} \hat{e}_{ij}\hat{e}_{ik}}{\sum_{i=1}^{K} \frac{1}{2} n_i(n_i - 1) - p}, \tag{2.20}$$

kde $\hat{\varphi}$ je odhad parametru φ, ktorý predstavíme neskôr.

Aj keď táto štruktúra predpokladá rovnakú koreláciu medzi pozorovaniami v skupine, ukazuje sa, že je vhodnou štruktúrou aj v prípade, ak sa korelácie medzi pozorovaniami v skupine líši.

Autoregresná korelácia

Táto korelačná štruktúra predpokladá určitú časovú závislosť usporiadaných pozorovaní. Uvažuje sa, že pozorovania v skupine tvoria autoregresný AR(m) proces. Pri tejto štruktúre je najťažšie určiť rád autoregresného procesu. Najbežnejšie používaným AR procesom v prípade korelačnej štruktúry je AR(1) proces. Prvky matice $\mathbb{R}_i(\boldsymbol{\alpha})$ sa riadia pravidlom

$$(\mathbb{R}_i)_{[k,l]} = \begin{cases} 1, & k = l \\ \alpha^{|k-l|}, & k \neq l, \end{cases}$$

kde $\alpha \in (0,1)$. Z pravidla vyššie vieme určiť tvar matice $\mathbb{R}_i(\boldsymbol{\alpha})$ a to

$$\mathbb{R}_i(\boldsymbol{\alpha}) = \begin{pmatrix} 1 & \alpha & \alpha^2 & \dots & \alpha^{n-1} \\ \alpha & 1 & \alpha & \dots & \alpha^{n-2} \\ \alpha^2 & \alpha & 1 & & \vdots \\ \vdots & & & \ddots & \vdots \\ \alpha^{n-1} & \dots & \dots & \alpha & 1 \end{pmatrix}. \tag{2.21}$$

V tomto prípade uvažujeme taktiež len odhad konštanty α. Ďalej uvažujeme, že pozorovania $Y_{i1}, \dots, Y_{in}$ tvoria AR(1) proces. Vieme, že $\mathbb{E}(e_{ij}e_{ik}) = \varphi\alpha^{|j-k|}$, teda parameter $\log \alpha$ môžeme odhadnúť ako smernicu regresného modelu, kde vysvetľovanou premennou bude $\log(\hat{e}_{ij}\hat{e}_{ik})$ a regresorom bude $|j - k|$.

Neštrukturovaná korelácia

Tuto necháme korelačnú štruktúru úplne nešpecifikovanú, čo z nej robí najkomplikovanejšiu, ale zároveň najflexibilnejšiu korelačnú štruktúru z pomedzi tých, ktoré sme už predstavili. Prvky matice sa budú riadiť pravidlom

$$(\mathbb{R}_i)_{[k,l]} = \begin{cases} 1, & k = l \\ \alpha_{kl}, & k \neq l. \end{cases}$$

Keďže sa jedná o korelačnú maticu, prvky na mieste $[k,l]$ budú rovné prvkom na mieste $[l,k]$ a matica $\mathbb{R}_i(\boldsymbol{\alpha})$ má tvar

$$\mathbb{R}_i(\boldsymbol{\alpha}) = \begin{pmatrix} 1 & \alpha_{12} & \alpha_{13} & \dots & \alpha_{1n} \\ \alpha_{12} & 1 & \alpha_{23} & \dots & \alpha_{2n} \\ \vdots & & \ddots & & \vdots \\ \vdots & & & \ddots & \alpha_{n-1n} \\ \alpha_{1n} & \alpha_{2n} & \dots & \alpha_{n-1n} & 1 \end{pmatrix}. \tag{2.22}$$

Kým v predchádzajúcich prípadoch bola matica $\mathbb{R}_i$ závislá len na konštante α, v tomto prípade bude závislá na vektore $\boldsymbol{\alpha} = (\alpha_{12},\alpha_{13},\ldots,\alpha_{1n},\alpha_{23},\ldots,\alpha_{n-1n})^\top$. To znamená, že bude potrebné odhadnúť $\frac{1}{2}n(n-1)$ korelácii medzi pozorovaniami. Preto sa táto štruktúra odporúča používať pre malé počty pozorovaní v skupine.

Odhad parametru $\tilde{\alpha}_{jk}$ pre koreláciu medzi pozorovaniami Y_{ij} a Y_{ik} má tvar

$$\tilde{\alpha}_{jk} = \frac{1}{\hat{\varphi}K}\sum_{i=1}^{K}\hat{e}_{ij}\hat{e}_{ik}, \quad j \neq k,$$

kde $\hat{\varphi}$ je opäť odhad parametru φ.

Fixná korelácia

Fixná korelačná štruktúra sa dá použiť v prípade, ak máme informáciu o danej korelačnej štruktúre z iných zdrojov, teda máme nejakú znalosť o koreláciách medzi pozorovaniami. V prípade použitia tejto štruktúry pri zovšeobecnených odhadovacích rovniciach nie je potreba štruktúru odhadovať, ale v priebehu odhadovacieho algoritmu ju uvažovať ako za pevne zvolenú korelačnú maticu. Ak parameter φ nie je neznámy, môžeme použiť metódu pseudo maximálnej vierohodnosti, ktorá bola predstavená v časti 2.2.

Poznámka. Výber korelačnej matice $\mathbb{R}(\boldsymbol{\alpha})$ je pre niektoré dátové štruktúry pomerne ťažký. Autori Hardin a Hilbe (2003) uvádzajú návod, ako vybrať vhodnú korelačnú štruktúru:

- pre balancované dáta, ktoré majú málo pozorovaní v skupine, sa odporúča voliť neštrukturovaná korelačná štruktúra,
- ak dáta v skupinách nie sú závislé na čase pozorovania, odporúča sa použiť spoločná korelačná štruktúra,
- ak sú data zoskupené a merané v závislosti na čase, mala by byť zvolená autoregresívna štruktúra,
- pre malý počet skupín použijeme nezávislú korelačnú štruktúru,
- pokiaľ dáta splňujú viac ako jednu špecifikáciu z vyššie uvedených, potom na výber vhodnej korelačnej štruktúry sa použije QIC kritérium (QIC kritérium bude predstavené v ďalšej kapitole)

Musíme ale upozorniť na fakt, že ak máme nejaké znalosti o skutočnej korelačnej štruktúre, mali by sme tieto znalosti použiť. V týchto prípadoch sa teda neodporúča ohliadať sa na návody uvedené vyššie.

Ďalšie korelačné štruktúry, ktoré sme neuviedli, môžeme nájsť v článku Liang a Zeger (1986) alebo v publikácii Hardin a Hilbe (2003).

2.4.2 Odhad parametru φ

V zovšeobecnených odhadovacích rovniciach podľa Liang a Zeger (1986) vieme explicitne vyjadriť odhad $\tilde{\varphi}$ parametru φ. V publikiáciách Liang a Zeger (1986) a Hardin a Hilbe (2003) sa uvádza najčastejšie používaný odhad φ , ktorý má tvar

$$\tilde{\varphi} = \frac{1}{N-p} \sum_{i=1}^{K} \sum_{j=1}^{n} \hat{e}_{ij}^2, \tag{2.23}$$

kde N je počet všetkých pozorovaní v skupinách a p je počet vysvetľujúcich premenných zaradených v modeli.

Niektoré zdroje uvádzajú odhad disperzného parametru taktiež ako

$$\tilde{\varphi} = \frac{1}{N} \sum_{i=1}^{K} \sum_{j=1}^{n} \hat{e}_{ij}^2. \tag{2.24}$$

Každý z týchto odhadov $\tilde{\varphi}$ parametru φ má svoje výhody. Podľa Hardin a Hilbe (2003), výhodou (2.23) je rovnaký odhad regresných koeficientov $\boldsymbol{\beta}$ ako v prípade GLM, ak za pracovnú korelačnú štruktúru zvolíme nezávislú korelačnú štruktúru. Na druhú stranu, použitím (2.24) dostávame rovnaký odhad regresných koeficientov $\boldsymbol{\beta}$ pri replikácii pozorovaní. Replikáciou pozorovaní rozumieme viac násobné nakopírovanie už sledovaných pozorovaní v datovom súbore. Týmto prístupom ostane odhad rovnaký, ale môžeme pozorovať zmenu smerodatnej chyby odhadu.

3. Praktická časť

V tejto kapitole sa budeme zameriavať na asymptotické vlastnosti odhadu $\hat{\boldsymbol{\beta}}$ parametru $\boldsymbol{\beta}$ na základe nami nasimulovaných dát. Generovanie dát a výpočty budú prebiehať v softvére R vo verzii 3.6.3.
V prvom prípade sa pozrieme, ako vplýva počet skupín K na vlastnosti odhadu $\hat{\boldsymbol{\beta}}$. V druhom prípade uvedieme príklad, ktorý má využitie v poisťovníctve. V ňom sa zameriame na správny výber korelačnej štruktúry pre skupinu pozorovaní.

3.1 Prvá simulačná štúdia

V simulačnej štúdii budeme sledovať, ako počet skupín K ovplyvňuje odhad $\hat{\boldsymbol{\beta}} = (\hat{\beta}_0,\hat{\beta}_1)^\top$. Budeme uvažovať model

$$Y_{ij} = \beta_0 + \beta_1 X_{ij} + \varepsilon_{ij}, \quad i = 1,\ldots,K, \quad j \leq n_i, \tag{3.1}$$

kde pre n_i platí

$$P(n_i = 2) = 0{,}1\,; P(n_i = 3) = 0{,}2\,; P(n_i = 4) = 0{,}3\,; P(n_i = 5) = 0{,}4,$$

pre počet skupín volíme $K = 100,\ 500,\ 1\,000$ a parameter $\boldsymbol{\beta} = (\beta_0,\beta_1)^\top$ položíme rovný $\beta_0 = 0{,}5$ a $\beta_1 = 0{,}3$. Vysvetľujúce premenné X_{ij} sú generované z alternatívneho rozdelenia $Alt(p)$ s $p = 0{,}5$. Pomocou funkcie rmvnorm z knižnice mvtnorm ďalej generujeme vektor chýb $\boldsymbol{\varepsilon}_i = (\varepsilon_{i1},\ldots,\varepsilon_{in_i})^\top$ z n_i-rozmerného normálneho rozdelenia so strednou hodnotou $\boldsymbol{\mu}_i = \mathbf{0}_{1\times n_i}$ a variančnou maticou

$$\boldsymbol{\Sigma}_i = \begin{pmatrix} 1 & \alpha & \ldots & \alpha \\ \alpha & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \alpha \\ \alpha & \ldots & \alpha & 1 \end{pmatrix}_{n_i\times n_i},$$

kde parameter α položíme rovný $\alpha = 0{,}8$.

Na základe vygenerovaných premenných X_{ij} a ε_{ij} podľa (3.1) dostávame pozorovania Y_{ij}, ktoré sú skupinovo závislé, ale nezávislé medzi skupinami. Z nagenerovaných dát vidíme, že skutočná korelačná štruktúra v skupine bude spoločná korelačná štruktúra. Pre každé K prevedieme 1 000 simulácií.

Pre každý vygenerovaný dataset odhadneme model a nájdeme odhad parametru $\boldsymbol{\beta}$. Ten budeme odhadovať funkciou geeglm z knižnice geepack s následovnými vstupmi:

1. Za štruktúru pracovnej korelačnej matice $\mathbb{R}(\alpha)$ budeme uvažovať spoločnú korelačnú štruktúru (EXCH).

2. Za štruktúru pracovnej korelačnej matice $\mathbb{R}(\alpha)$ budeme uvažovať autoregresnú AR(1) štruktúru (AR).

3. Za štruktúru pracovnej korelačnej matice $\mathbb{R}(\alpha)$ budeme uvažovať neštrukturovanú korelačnú štruktúru (UNS).

4. Za štruktúru pracovnej korelačnej matice $\mathbb{R}(\alpha)$ budeme uvažovať nezávislú korelačnú štruktúru (IND).

Ďalej odhadneme model funkciou glm z balíčku stats. Tento prípad budeme uvažovať z toho dôvodu, že obdržíme rovnaké odhady parametru $\boldsymbol{\beta}$ ako v prípade voľby nezávislej korelačnej štruktúry. Rozdiel medzi týmito dvoma prístupmi je ten, že v prípade odhadu GLM (ako je uvedené v kapitole 1) zanedbávame rozdelenie pozorovaní Y_{ij} do skupín pre odhad rozptylu parametru $\boldsymbol{\beta}$.

Primárne sa zameriame na odhad $\hat{\beta}_1$ parametru β_1. V tabuľke 3.1 môžeme vidieť percentuálne pokrytie skutočnej hodnoty parametru β_1 konfidenčným 95%-percentným intervalom. Vidíme, že pre metódu GEE s vyššie uvedenými korelačnými štruktúrami a pre všetky hodnoty K je pokrytie skutočného parametru dostatočne vysoké. Môžeme si všimnúť, že pokrytie pre odhad parametru pomocou metódy GLM je výrazne nižšie oproti ostatným hodnotám pokrytia.

K	**EXCH**	**AR**	**UNS**	**IND**	**GLM**
100	95,2%	95,2%	93,1%	94,6%	69,5%
500	94,7%	94,2%	94,3%	94,5%	69,3%
1000	94,4%	94,0%	94,4%	94,5%	69,2%

Tabuľka 3.1: Pokrytie skutočnej hodnoty parametru β_1 95% konfidenčným intervalom.

V tabuľke 3.2 sú uvedené priemerné dĺžky konfidenčných intervalov. 95% konfidenčný interval má tvar

$$\left(\hat{\beta}_1 \pm u_{0,975} \cdot \sqrt{\widehat{\text{var}}(\hat{\beta}_1)}\right), \tag{3.2}$$

kde $u_{0,975} = 1.96$ je kvantil normálneho rozdelenia $N(0,1)$ a $\widehat{\text{var}}(\hat{\beta}_1)$ je odhadnutý rozptyl odhadu $\hat{\beta}_1$. Môžeme si všimnúť, že so zväčšujúcim sa počtom skupín K je priemerná dĺžka intervalov menšia. Povšimnime si, že pre počet skupín $K = 100$ a pre odhad metódou GEE s neštrukturovanou korelačnou štruktúrou je priemerná dĺžka konfidenčného intervalu skoro dvojnásobne vyššia, ako je to pri ostatných korelačných štruktúrach. Usudzujeme, že je to spôsobené tým, že v tejto štruktúre je potreba odhadnúť viac parametrov a do odhadu rozptylu vstupuje ďalšia variabilita.

K	**EXCH**	**AR**	**UNS**	**IND**	**GLM**
100	0,718	0,731	1,331	0,736	0,391
500	0,323	0,329	0,382	0,331	0,175
1000	0,229	0,233	0,229	0,235	0,124

Tabuľka 3.2: Priemerná dĺžka 95% intervalov spoľahlivosti.

Pozrime sa taktiež na rozdiel medzi odhadom GEE s nezávislou korelačnou štruktúrou a odhadom GLM. Ako môžeme vidieť v tabuľke 3.2, priemerné dĺžky in-

tervalov pre metódu GEE s nezávislou korelačnou štruktúrou sú približne dvojnásobne dlhšie ako dĺžky intervalov pre metódu GLM. Odtiaľ vidíme, že ak zanedbáme rozdelenie pozorovaní do skupín, dostaneme metódou GLM odhad s podhodnoteným odhadnutým rozptylom a teda konfidenčný interval nepokryje skutočnú hodnotu parametru β_1 s predpísanou pravdepodobnosťou.

Ďalej nahliadnime na odhad strednej štvorcovej chyby odhadu $\hat{\beta}_1$. Odhad strednej štvorcovej chyby má tvar

$$\widehat{MSE}(\hat{\beta}_1) = \frac{1}{B}\sum_{b=1}^{B}(\hat{\beta}_1^{(b)} - \beta_1^{(b)})^2,$$

kde $B = 1\,000$ je počet simulácií. Tento odhad strednej štvorcovej chyby nám vyjadruje presnosť odhadu ako priemernú hodnotu druhej mocniny rozdielu skutočného parametru β_1 a odhadu $\hat{\beta}_1$. Čím menšia je táto hodnota, tým je odhad $\hat{\beta}_1$ "bližšie" ku skutočnému parametru β. V tabuľke 3.3 vidíme hodnoty 100-násobku odhadu strednej štvorcovej chyby pre nami odhadnuté modely.

K	**EXCH**	**AR**	**UNS**	**IND**	**GLM**
100	3,361	3,510	145,403	3,572	3,572
500	0,714	0,740	0,771	0,748	0,748
1000	0,341	0,360	0,357	0,355	0,355

Tabuľka 3.3: Hodnoty 100-násobku odhadu strednej štvorcovej chyby odhadu $\hat{\beta}_1$ parametru β_1.

Ako môžeme vidieť, najnižšie hodnoty odhadu strednej štvorcovej chyby dostávame pri modeli odhadnutom metódou GEE so spoločnou korelačnou štruktúrou. To sme aj očakávali, keďže táto korelačná štruktúra bola skutočnou štruktúrou pri generovaní dát pre simuláciu.
Taktiež vidíme, že hodnoty odhadu strednej štvorcovej chyby pre metódu GEE s nezávislou korelačnou štruktúrou a pre metódu GLM rovnaké. Ako sme už spomínali vyššie, je to tým, že oboma týmito metódami dostávame rovnaké hodnoty odhadov $\hat{\beta}_1$
Pre $K = 100$ a pre neštrukturovanú korelačnú štruktúru môžeme pozorovať vyššiu hodnotu odhadnutej strednej štvrocovej chyby. To môže byť spôsobené tým, že nemáme k dispozícii dostatočné množstvo pozorovaní a preto je odhad $\hat{\beta}_1$ od hodnoty skutočného parametru β_1 "ďalej".

3.2 Druhá simulačná štúdia

V tejto časi práce sa zamierame na využitie GEE v praxi. Budeme pozorovať počet škodových udalostí motorových vozidiel klientov poisťovne. Budeme uvažovať $K = 3\,000$ klientov, kde dĺžka obdobia poistenia, ktoré má klient zjednané, je v rozmedzí dvoch až štyroch rokov.

Na začiatok popíšeme, ako sú generované data pre simuláciu. Pre každého klienta sme najprv vygenerovali poistné obdobie, počas ktorého sme pozorovali počet škodových udalostí. To nám udáva počet pozorovaní n_i, $i = 1, \ldots, K$ pre klienta (teda počet pozorovaní v skupine). Predpokladáme, že pravdepodobnosť $n_i = 2$ je rovná 0,1, pravdepodobnosť $n_i = 3$ je rovná 0,4 a pravdepodobnosť $n_i = 4$ je rovná 0,5. Ako vysvetľujúce premenné budeme uvažovať vek klienta, pohlavie klienta a to, aké motorové vozidlo klient vlastní.

Pre pohlavie budeme uvažovať hodnotu 0, ak sa jedná o muža a hodnotu 1, ak sa jedná o ženu. Budeme predpokladať, že obe hodnoty budú zastúpené rovnako, teda pravdepodobnosť výskytu mužov a žien bude 0,5. Pre veličinu veku budeme uvažovať hodnoty od 18 do 70 rokov. Tu budeme taktiež predpokladať, že výskyt hodnôt v tomto rozmedzí bude zastúpený rovnako a to s pravdepodobnosťou $\frac{1}{52}$. Pre motorové vozidlo budeme uvažovať tri hodnoty- osobný automobil (A), motorku (M) a dodávku (D). Predpokladáme, že pravdepodobnosť, že klient vlastní osobný automobil je rovná 0,6, pravdepodobnosť, že vlastní motorku je rovná 0,3 a pravdepodobnosť, že vlastní dodávku je rovná 0,1. Nakoniec budeme predpokladať, že počas pozorovaného obdobia nedošlo k úmrtiu žiadneho klienta a taktiež, že nenastala zmena vo vlastníctve motorového vozidla.

Budeme uvažovať model

$$\mathbb{E}\left[Y_{ij}|X_{ij}\right] = \exp\left\{\beta_0 + \beta_1 \cdot pohlavie + \beta_2 \cdot (vek - 18) + \beta_3 \cdot \mathbb{I}_M + \beta_4 \cdot \mathbb{I}_D\right\}, \tag{3.3}$$

kde Y_{ij} je počet škôd i-teho klienta v j-tom roku poistného obdobia, $i = 1, \ldots, K$, $j \leq n_i$, $\mathbb{I}_M$ je indikátor, že klient vlastní motorku a $\mathbb{I}_D$ je indikátor, že klient vlastní dodávku. Teraz prejdime ku generovaniu počtu škodových udalostí $\boldsymbol{Y}_i$. Označme

$$\mathbb{E}\left[\boldsymbol{Y}_i|\mathbb{X}_i\right] = \boldsymbol{\lambda}_i = \exp\{\mathbb{X}_i\boldsymbol{\beta}\}, \tag{3.4}$$

kde $\mathbb{X}_i$ je regresná matica pre i-teho klienta a $\boldsymbol{\beta}$ je regresný parameter. Kedže sa jedná o počet škodových udalostí, hodnoty $\boldsymbol{Y}_i$ budú generované z Poissonovho rozdelenia

$$\boldsymbol{Y}_i \sim Pois(\boldsymbol{\lambda}_i),$$

kde $\boldsymbol{\lambda}_i = (\lambda_{i1}, \ldots, \lambda_{in_i})$ a $Y_{ij} \sim Pois(\lambda_{ij})$. Ak by sme ale generovali hodnoty $\boldsymbol{Y}_i$ z Poissonovho rozdelenia s parametrom $\boldsymbol{\lambda}_i$, dostali by sme navzájom nezávislé pozorovania. Preto parameter $\boldsymbol{\lambda}_i$ upravíme tak, aby pozorovania v i-tej skupine boli navzájom závislé, ale nezávislé medzi skupinami. Označme

$$\boldsymbol{\lambda}_i^* = \exp\{\mathbb{X}_i\boldsymbol{\beta}\}Z_i,$$

kde $\boldsymbol{\lambda}_i^* = (\lambda_{i1}^*, \ldots, \lambda_{in_i}^*)$ a platí, že Z_i sú generované nezávisle z rozdelenia s jednotkovou strednou hodnotou $\mathbb{E}Z_i = 1$. Pre účely našej simulácie sme hodnoty Z_i generovali z exponenciálneho rozdelenia Exp(1). Na základe tejto úpravy budeme generovať $\boldsymbol{Y}_i$ z Poissonovho rozdelenia s parametrom $\boldsymbol{\lambda}_i^*$. Ukážeme, že napriek tejto úprave bude platiť rovnosť v (3.4):

$$\begin{aligned}\mathbb{E}\left[\boldsymbol{Y}_i|\mathbb{X}_i\right] = \mathbb{E}\left[\mathbb{E}\left[\boldsymbol{Y}_i|\mathbb{X}_i, Z_i\right]|\mathbb{X}_i\right] &= \mathbb{E}\left[\exp\{\mathbb{X}_i\boldsymbol{\beta}\}Z_i|\mathbb{X}_i\right] \\ &= \exp\{\mathbb{X}_i\boldsymbol{\beta}\}\mathbb{E}\left[Z_i|\mathbb{X}_i\right] = \exp\{\mathbb{X}_i\boldsymbol{\beta}\}.\end{aligned}$$

Keďže vieme, z akého rozdelenia sú generované $\boldsymbol{Y}_i$, môžeme zistiť, ako vyzerá skutočná kovariančná matica $\boldsymbol{Y}_i$ za podmienky $\mathbb{X}_i$:

$$\text{var}\left[\boldsymbol{Y}_i|\mathbb{X}_i\right] = \mathbb{E}\left[\text{var}\left[\boldsymbol{Y}_i|\mathbb{X}_i,Z_i\right]|\mathbb{X}_i\right] + \text{var}\left[\mathbb{E}\left[\boldsymbol{Y}_i|\mathbb{X}_i,Z_i\right]|\mathbb{X}_i\right] \tag{3.5}$$

Prvý člen v (3.5) môžeme rozpísať ako

$$\mathbb{E}\left[\text{var}\left[\boldsymbol{Y}_i|\mathbb{X}_i,Z_i\right]|\mathbb{X}_i\right] = \mathbb{E}\left[\boldsymbol{\Sigma}_i^*|\mathbb{X}_i\right],$$

kde $\boldsymbol{\Sigma}_i^*$ je skutočná kovariančná matica $\boldsymbol{Y}_i$ za podmienky $\mathbb{X}_i$ a Z_i. Tú vieme vyjadriť ako

$$\boldsymbol{\Sigma}_i^* = \begin{pmatrix} \lambda_{i1}^* & \cdots & \rho_{1n_i}^* \\ \vdots & \ddots & \vdots \\ \rho_{1n_i}^* & \cdots & \lambda_{in_i}^* \end{pmatrix},$$

kde na diagonále sú rozptyly Y_{ij}, ktoré poznáme na základe znalosti rozdelenia Y_{ij}, a mimo diagonály sú korelácie medzi pozorovaniami, to znamená $\rho_{1n_i}^*$ je korelácia medzi Y_{i1} a Y_{in_i}. Koreláciu medzi Y_{ij} a Y_{ik} vieme spočítať ako

$$corr(Y_{ij},Y_{ik}|\mathbb{X}_i,Z_i) = \rho_{jk}^* = \frac{\mathbb{E}\left[(Y_{ij}-\lambda_{ij}^*)(Y_{ik}-\lambda_{ik}^*)\right]}{\lambda_{ij}^*\lambda_{ik}^*} = 0, \tag{3.6}$$

keďže pozorovania Y_{ij} a Y_{ik} sú za podmienky $\mathbb{X}_i$ a Z_i nezávislé.

Druhý člen v (3.5) môžeme písať ako

$$\begin{aligned}\text{var}\left[\mathbb{E}\left[\boldsymbol{Y}_i|\mathbb{X}_i,Z_i\right]|\mathbb{X}_i\right] = \text{var}\left[\exp\{\mathbb{X}_i\boldsymbol{\beta}\}Z_i|\mathbb{X}_i\right] &= \exp\{\mathbb{X}_i\boldsymbol{\beta}\}\,\text{var}\left[Z_i|\mathbb{X}_i\right]\exp\{\mathbb{X}_i\boldsymbol{\beta}\}^\top \\ &= \exp\{\mathbb{X}_i\boldsymbol{\beta}\}\exp\{\mathbb{X}_i\boldsymbol{\beta}\}^\top.\end{aligned}$$

Na základe vyššie popísaného generovania dát prevedieme 1 000 simulácií. Parameter regresných koeficientov $\boldsymbol{\beta}$ budeme voliť ako

$$\boldsymbol{\beta} = (\beta_0,\beta_1,\beta_2,\beta_3,\beta_4)^\top = (0{,}8\,;\,0{,}5\,;-0{,}1\,;\,0{,}4\,;-0{,}3)^\top.$$

Pre každú simuláciu odhadneme model uvedený v (3.3) funkciou gee z knižnice gee s nasledujúcimi vstupmi:

1. Za štruktúru pracovnej korelačnej matice $\mathbb{R}(\alpha)$ budeme uvažovať spoločnú korelačnú štruktúru (EXCH).
2. Za štruktúru pracovnej korelačnej matice $\mathbb{R}(\alpha)$ budeme uvažovať autoregresnú AR(1) štruktúru (AR).
3. Za štruktúru pracovnej korelačnej matice $\mathbb{R}(\alpha)$ budeme uvažovať nezávislú korelačnú štruktúru (IND).
4. Za štruktúru pracovnej korelačnej matice $\mathbb{R}(\alpha)$ budeme uvažovať neštrukturovanú korelačnú štruktúru (UNS).

Pre každú z možností uvedených vyššie budeme uvažomať maximálny počet iterácii 10.

Teraz prejdime k definícii QIC kritéria, ktoré sa používa k nájdeniu vhodnej korelačnej štruktúry.

3.2.1 QIC kritérium

V článku Pan (2001) sa pre určenie vhodnej korelačnej štruktúry, kde sa rozhodujeme medzi viacerými korelačnými štruktúrami, odporúča použiť QIC kritérium. Toto kritérium je odvodené od Akaikeho informačného kritéria (AIC). Narozdiel od AIC, ktoré využíva metódu vierohodnosti, je QIC založené na kvázi vierohodnosti. Taktiež využíva predpoklad nezávislosti pozorovaní. Odtiaľ vyplýva aj názov tohto kritéria - kvázi-vierohodnostné informačné kritérium za predpokladu nezávislosti. Kvázi vierohodnostná funkcia pre Y_{ij}, ako ju uvádzajú McCullagh a Nelder (1989) má tvar

$$Q(\mu_{ij},\varphi,Y_{ij}) = \int_{Y_{ij}}^{\mu_{ij}} \frac{Y_{ij}-t}{\varphi V(t)} dt,$$

kde φ je disperzný parameter rozdelenia z exponenciálnej rodiny rozdelení uvedenej v časti 1.1. Keďže sa predpokladá nezávislosť pozorovaní, dostávame kvázi vierohodnostnú funkciu pre $\boldsymbol{Y}$

$$Q(\boldsymbol{\mu},\varphi,\boldsymbol{Y}) = \sum_{i=1}^{K}\sum_{j=1}^{n_i} Q(\mu_{ij},\varphi,Y_{ij}),$$

kde $\boldsymbol{\mu} = (\boldsymbol{\mu}_1,\ldots,\boldsymbol{\mu}_K)^\top$ a $\boldsymbol{Y} = (\boldsymbol{Y}_1,\ldots,\boldsymbol{Y}_K)^\top$.

QIC kritérium je v článku Pan (2001) definované ako

$$QIC = -2Q(\boldsymbol{Y},\varphi,\hat{\boldsymbol{\mu}}_{\mathbb{R}}) + 2\text{trace}(\boldsymbol{\Omega}^{-1}\hat{\mathbb{V}}_{\mathbb{R}}),$$

kde $\hat{\boldsymbol{\mu}}_{\mathbb{R}} = g^{-1}(\mathbb{X}\hat{\boldsymbol{\beta}}_{\mathbb{R}})$, $\mathbb{X}$ je regresná matica a $\hat{\boldsymbol{\beta}}_{\mathbb{R}}$ je odhad regresných koeficientov pri predpokladanej korelačnej štruktúre $\mathbb{R}$. Ďalej $\boldsymbol{\Omega}$ je odhad kovariančnej matice odhadu $\hat{\boldsymbol{\beta}}$ parametru $\boldsymbol{\beta}$ získaná z modelu s nezávislou štruktúrou a $\hat{\mathbb{V}}_{\mathbb{R}}$ je sendvičový odhad kovariančnej matice (ako je uvedené v bode 3 na strane 22 tejto práce) odhadu $\hat{\boldsymbol{\beta}}$ parametru $\boldsymbol{\beta}$ pri predpokladanej korelačnej štruktúre $\mathbb{R}$. Pre výber najvhodnejšej korelačnej štruktúry vyberieme tú, pre ktorú je hodnota QIC kritéria najmenšia.

Pre každý model sme vypočítali hodnotu QIC kritéria pomocou funkcie QIC z knižne gee a zaznamenali, pre ktorú korelačnú štruktúru bola hodnota QIC kritéria najnižšia. V tabuľke 3.4 môžeme vidieť, v percentuálne koľkých prípadoch QIC kritérium vybralo danú korelačnú štruktúru.

EXCH	**AR**	**UNS**	**IND**
70,1%	3,9%	18,7%	7,3%

Tabuľka 3.4: Percentuálne zastúpenie výberu QIC kritéria.

Vidíme, že QIC kritérium zvolilo za vhodnú korelačnú štruktúru spoločnú korelačnú štruktúru až v 70% prípadov. Z toho môžeme usúdiť, že táto korelačná štruktúra by mohla byť skutočnou korelačnou štruktúrou. Na základe vyššie opísaného QIC kritéria vyberieme v každej simulácii taký odhad, ktorý odpovedá

korelačnej štruktúre vybranej týmto kritériom.

V tabuľke 3.5 môžeme vidieť percentuálne pokrytie skutočnej hodnoty zložiek parametru $\boldsymbol{\beta}$. Vidíme, že medzi odhadmi GEE s korelačnými štruktúrami nie je výrazný rozdiel v pokrytí hodnoty zložiek skutočného parametru $\boldsymbol{\beta}$ a toto pokrytie pre každú zložku je aspoň 94,0%. Rozdiel opäť, ako v prvej simulačnej štúdii, vidíme pri pokrytí odhadov metódou GLM.

	EXCH	**AR**	**UNS**	**QIC**	**IND**	**GLM**
β_0	95,7%	95,8%	95,5%	95,5%	95,9%	61,9%
β_1	95,2%	95,4%	95,3%	95,2%	95,6%	58,0 %
β_2	95,3%	94,7%	94,8%	94,9%	94,6%	66,6%
β_3	94,5%	94,3%	94,4%	94,6%	94,4%	53,6%
β_4	94,4%	93,9%	94,0%	94,2%	94,3%	64,6%

Tabuľka 3.5: Pokrytie skutočnej hodnoty parametru $\boldsymbol{\beta}$ konfidenčným 95% intervalom.

Ďalej môžeme v tabuľke 3.6 vidieť priemernú dĺžku 95% konfidenčných intervalov. Môžeme pozorovať, že priemerná dĺžka konfidenčného intervalu pre metódu GEE je najmenšia pre odhady zložiek parametru $\boldsymbol{\beta}$, kde sme korelačnú štruktúru zvolili na základe QIC kritéria. Ďalej vidíme, že zo štyroch korelačných štruktúr dostávame najkratšiu priemernú dĺžku intervalu pre spoločnú korelačnú štruktúru.

	EXCH	**AR**	**UNS**	**QIC**	**IND**	**GLM**
β_0	0,266	0,268	0,267	0,265	0,269	0,111
β_1	0,262	0,263	0,262	0,261	0,264	0,105
β_2	0,110	0,112	0,111	0,110	0,112	0,054
β_3	0,289	0,290	0,289	0,288	0,291	0,107
β_4	0,456	0,459	0,457	0,455	0,459	0,212

Tabuľka 3.6: Priemerná dĺžka 95% konfidenčného intervalu. Pre β_2 uvádzame 10-násobok priemernej dĺžky 95% konfidenčného intervalu.

Nakoniec v tabuľke 3.7 vidíme odhad strednej štvrocovej chyby

$$\widehat{MSE}(\hat{\beta}_i) = \frac{1}{B}\sum_{b=1}^{B}(\hat{\beta}_i^{(b)} - \beta_i^{(b)})^2,$$

kde $B = 1\,000$ je počet prevedených simulácií a $\hat{\beta}_i^{(b)}$ je odhad parametru β_i v b-tej simulácii. Pozorujeme, že odhad strednej štvorcovej chyby sa medzi korelačnými štruktúrami výrazne nelíši. Najnižšie hodnoty odhadu strednej štvrocovej chyby dostávame pre odhady metódou GEE so spoločnou korelačnou štruktúrou a s korelačnou štruktúrou zvolenou na základe QIC kritéria. Môžeme si všimnúť, že pre odhad parametru β_4 dostávame vyššiu hodnotu odhadu strednej štvrocovej chyby, ako pri ostatných parametroch. Predpokladáme, že to je spôsobené menším výskytom klientov, ktorý vlastnia dodávku.

Opäť si môžeme všimnúť, že hodnoty odhadnutej strednej štvorcovej chyby pre metódu GEE s nezávislou korelačnou štruktúrou a metódy GLM sú rovnaké. Ako sme už vysvetlili v prvej simulačnej štúdii, je to spôsobené rovnakými hodnotami odhadov zložiek parametru $\boldsymbol{\beta}$.

	EXCH	**AX**	**UNS**	**QIC**	**IND**	**GLM**
β_0	0,423	0,427	0,426	0,425	0,432	0,432
β_1	0,424	0,427	0,427	0,425	0,432	0,432
β_2	0,080	0,082	0,081	0,080	0,083	0,083
β_3	0,562	0,567	0,560	0,561	0,563	0,563
β_4	1,369	1,401	1,387	1,367	1,403	1,403

Tabuľka 3.7: Hodnoty 100-násobku odhadu strednej štvrocovej chyby odhadu $\hat{\boldsymbol{\beta}}$ parametru $\boldsymbol{\beta}$. Pre β_2 uvádzame hodnotu 10 000-násobku odhadu strednej štvorcovej chyby.

Na základe výsledkov opísaných vyššie vidíme, že z výberu štyroch korelačných štruktúr dostávame najlepšie výsledky pre spoločnú korelačnú štruktúru. Ďalej z výsledkov vidíme, že pre voľbu korelačnej štruktúry je možné využiť QIC kritérium, pokiaľ sa nevieme rozhodnúť, ktorú korelačnú štruktúru zvoliť. Usudzujeme to na základe toho, že pre spoločnú korelačnú štruktúru dostávame najlepšie výsledky a práve túto štruktúru zvolilo QIC kritérium v najviac prípadoch.

Záver

V tejto práci sme sa venovali metódam zovšeobecnených odhadovacích rovníc (GEE). Na začiatok sme predstavili rodinu exponenciálnych rozdelení a ukázali sme vlastnosti rozdelení, ktoré do nej patria. Formulovali sme zovšeobecnený lineárny model a hľadali odhady jeho parametrov metódou maximálnej vierohodnosti.

V druhej časti práce sme pristúpili k zovšeobecneným odhadovacím rovniciam. Predstavili sme metódy, za pomoci ktorých nájdeme konzistentné odhady parametru. Prvá metóda, s ktorou sme čitateľa oboznámili je metóda pseudo maximálnej vierohodnosti. Odvodili sme tvar odhadovacej rovnice a predstavili asymptotické výsledky odhadu parametru. Obdobne sme postupovali aj pri druhej metóde, pri metóde kvázi psuedo maximálnej vierohodnosti. Ďalej sme sa zamerali na metódu zovšeobecnených rovníc (GEE) podľa Liang a Zeger (1986). Popísali sme iteratívny algoritmus, ako nájsť odhad parametru pomocou tejto metódy. Nakoniec sme predstavili súbor najčastejšie používaných korelačných štruktúr.

V poslednej časti práce sme sa zamerali na simulačné štúdie. V prvej simulačnej štúdii sme ukázali, ako počet skupín ovplyvňuje vlastnosti odhadu parametru. Ukázali sme, že čím väčší počet skupín pozorovaní máme, tým viac dostávame lepšie vlastnosti odhadu parametru. V druhej simulačnej štúdii sme sa zamerali na príklad z praxe. Ukázali sme, že QIC kritérium je naozaj vhodným informačným kritériom pre výber vhodnej pracovnej korelačnej štruktúry, pokiaľ sa nevieme rozhodnúť medzi štruktúrami, ktoré by mohli byť skutočné.

Zoznam použitej literatúry

Fitzmaurice, G. M., Laird, N. M. a Ware, J. H. (2004). *Applied Longitudinal Analysis*. John Wiley & Sons, Hoboken, New Jersey. ISBN 978-0471214878.

Gourieroux, C., Monfort, A. a Trognon, A. (1984). Pseudo maximum likelihood methods: Theory. *Econometrica*, **52**(3), 681–700. doi: 10.2307/1913471.

Hardin, J. W. a Hilbe, J. M. (2003). *Generalized estimating equations*. Chapman & Hall/CRC, Boca Raton, Florida. ISBN 1584883073.

Jiang, J. (2010). *Large Sample Techniques for Statistics*. Springer New York, New York. ISBN 978-1-419-6826-5.

Kariya, T. (2004). *Generalized Least Squares*. John Wiley & Sons, Hoboken, New Jersey. ISBN 0-470-86697-7.

Kulich, M. (2020). Course notes to advanced regression models. URL `http://www.karlin.mff.cuni.cz/~kulich/vyuka/pokreg/doc/advreg_notes_200522.pdf`. dátum prístupu: 28.05.2020.

Lachout, P. (2004). *Teorie pravděpodobnosti*. Nakladatelství Karolinum, Praha. ISBN 80-246-0872-3.

Liang, K.-Y. a Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, **73**(1), 13–22. doi: 10.2307/2336267.

McCullagh, P. a Nelder, J. A. (1989). *Generalized Linear Models*. Chapman & Hall/CRC, Boca Raton, Florida. ISBN 978-0-412-31760-6.

Nelder, J. A. a Wedderburn, R. W. M. (1972). Generalized linear models. *Journal of the Royal Statistical Society*, **135**(3), 370–384. doi: 10.2307/2344614.

Pan, W. (2001). Model selection in estimating equations. *Biometrics*, **57**(2), 529–534. doi: 10.1111/j.0006-341x.2001.00529.x.

White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica*, **50**(1), 1–25. doi: 10.2307/1912526.

Ziegler, A. (2011). *Generalized Estimating Equations*. Springer New York, New York. ISBN 978-1-4614-0498-9.

Zvára, K. a Štepán, J. (1997). *Pravděpodobnost a matematická statistika*. Matfyzpress, Praha. ISBN 80-85863-24-3.

A. Prílohy

A.1 Vlastnosti n-dimenzionálnej exponenciálnej rodiny rozdelení

V tejto časti uvedieme tvrdenia, ktoré sme využívali primárne v druhej kapitole tejto práce. Jedná sa o vlastnosti mnohorozmernej exponenciálnej rodiny rozdelení.

Tvrdenie 2. *(Ziegler (2011), Property 1.3) Pre rozdelenia patriace do mnohorozmernej rodiny exponenciálnych rozdelení platí*

$$\boldsymbol{\Sigma}^{-1} = \frac{\partial a(\boldsymbol{\mu},\boldsymbol{\Phi})}{\partial \boldsymbol{\mu}^\top} = \boldsymbol{V}(\boldsymbol{\mu},\boldsymbol{\Phi}),$$

kde $\boldsymbol{V}(.,.)$ je takzvaná variančná funkcia.

Tvrdenie 3. *(Ziegler (2011), Property 1.4) Pre rozdelenia patriace do mnohorozmernej rodiny exponenciálnych rozdelení platí*

$$\left(\frac{\partial d(\boldsymbol{\mu},\boldsymbol{\Phi})}{\partial \boldsymbol{\mu}} + \frac{\partial a(\boldsymbol{\mu},\boldsymbol{\Phi})^\top}{\partial \boldsymbol{\mu}}\boldsymbol{Y}\right) = \left(-\frac{\partial a(\boldsymbol{\mu},\boldsymbol{\Phi})^\top}{\partial \boldsymbol{\mu}}\boldsymbol{\mu} + \frac{\partial a(\boldsymbol{\mu},\boldsymbol{\Phi})^\top}{\partial \boldsymbol{\mu}}\boldsymbol{Y}\right).$$

A.2 Definícia $K^{1/2}$-konzistencie

Definícia 2. *(Lachout (2004), Definice 6.2) Nech $a_n > 0$, $n \in \mathbb{N}$ je postupnosť čísel a X_n, $n \in \mathbb{N}$ je postupnosť náhodných veličín.*

1. *Povieme, že $X_n = o_p(a_n)$, ak pre každé $\epsilon > 0$ je*

$$\lim_{n\to\infty} P(|X_n| > \epsilon a_n) = 0.$$

2. *Povieme, že $X_n = O_p(a_n)$, ak*

$$\lim_{K\to\infty} \sup_{n\in\mathbb{N}} P(|X_n| > K a_n) = 0.$$

Definíciu 2 pre $o_P(a_n)$ a $O_P(a_n)$ môžeme rozšíriť aj pre postupnosť náhodných vektorov $\boldsymbol{X}_n$, $n \in \mathbb{N}$, ako uvádza Jiang (2010), strana 59.

Definícia 3. *Nech $a_n > 0$, $n \in \mathbb{N}$ je postupnosť čísel, $\boldsymbol{X}_n$, $n \in \mathbb{N}$ je postupnosť náhodných vektorov a $||.||$ je Euklidovská norma.*

1. *Povieme, že $\boldsymbol{X}_n = o_p(a_n)$, ak $||\boldsymbol{X}_n|| = o_p(a_n)$.*
2. *Povieme, že $\boldsymbol{X}_n = O_p(a_n)$, ak $||\boldsymbol{X}_n|| = O_p(a_n)$.*

Nech $\hat{\boldsymbol{\theta}}_n$ je vektorový odhad skutočného vektorového parametra $\boldsymbol{\theta}$ odhadnutý z pozorovaní $\boldsymbol{X} = (X_1, \ldots, X_n)^\top$. Ďalej nech postupnosť $\{a_n\}$ je postupnosť kladných čísel divergujúcich do nekonečna. Povieme, že $\hat{\boldsymbol{\theta}}_n$ je *a_n-konzistentný* odhad parametru $\boldsymbol{\theta}$ práve vtedy, keď

$$a_n(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}) = O_p(1).$$